2016-04-24

A Comprehensive Comparative Performance Evaluation of Signal Processing Features in Detecting Alcohol Consumption from Gait Data

Muxi Qi
Worcester Polytechnic Institute

A Comprehensive Performance Comparison of Signal Processing Features in Detecting Alcohol Consumption from Gait Data

A Thesis

Submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Electrical and Computer Engineering

By

Muxi Qi

Date: April, 2016

APPROVED:

Professor Emmanuel Agu, Thesis Adviser

Professor Kaveh Pahlavan, Thesis Committee Member

Professor Xinming Huang, Thesis Committee Member

Abstract

Excessive alcohol is the third leading lifestyle-related cause of death in the United States. Alcohol intoxication has a significant effect on how the human body operates, and is especially harmful to the human brain and heart. To help individuals to monitor their alcohol intoxication, several methods have been proposed to detect alcohol consumption levels including direct Blood Alcohol Concentration (BAC) measurement by breathalyzers and various wearable sensor devices. More recently, Arnold *et al* proposed a machine-learning-based method of passively inferring intoxication levels from gait data by classifying smartphone accelerometer readings. Their work utilized 11 smartphone accelerometer features in the time and frequency domains, achieving a classification accuracy of 57%.

This thesis extends the work of Arnold *et al* by extracting and comparing the efficacy of a more comprehensive list of 27 signal processing features in the time, frequency, wavelet, statistical and information theory domains, evaluating how much using them improves the accuracy of supervised BAC classification of accelerometer gait data. Correlation-based Feature Selection (CFS) is used to identify and rank features most correlated with alcohol-induced gait changes. 22 of the 27 features investigated showed statistically significant correlations with BAC levels. The most correlated features were then used to classify labeled samples of intoxicated gait data in order to test their detection accuracy. Statistical features had the best classification accuracy of 83.89%, followed by time domain features and frequency domain features follow with accuracies of 83.22% and 82.21%, respectively. Classification using all 22 statistically significant signal processing features yielded an accuracy of 84.9% for the Random Forest classifier.

Keywords: alcohol consumption, gait, smartphone, signal processing, machine learning

Acknowledgments

I would like to thank my adviser, Professor Emmanuel Agu, for his helpful guidance, and answering plenty of questions over the last 2 semesters during my research. His introduction of cutting edge techniques in gait analysis brought me into an interesting and practical field where I could apply what I have learnt in signal processing. This field is also related to so many other fields that attract me a lot. His guidance helped me avoid incorrect steps and time-consuming mistakes and always left enough room for me to try things by myself and improve my ability in research.

I would also like to thank Christina Aiello for her help in collecting the gait data I analyzed and introducing me to machine learning classification in the WEKA data mining library. Without the huge amount of data she collected and provided to me, I would not have been able to apply and test my ideas in a real environment. And also, her introduction of WEKA classification helped me to test the accuracy of the grouped features, because I was previously not familiar with WEKA tools.

Lastly, a great thanks to Professors Kaveh Pahlavan and Xinming Huang, for agreeing to be on my thesis committee. I would like to thank them for helpful discussions and questions that enabled me to look further and improve this thesis.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

1.1 Historical Background of Alcohol Consumption and Detection

Drinking is somehow part of our daily lives, for pleasure and business. In 2013, 56.4 percent of people ages 18 or older reported that they drank in the past month [23], 24.6 percent of people aged 18 or older reported that they engaged in binge drinking[1] in the past month, 8.6 percent reported heavy drinking [24].

Overdrinking may cause many problems, from individual to the society and is harmful to the human body. 46.4 percent of 71713 total liver disease deaths among people aged 12 and older involved alcohol in 2013 [25]. Overdrinking also increases the risk of cancers of the mouth, esophagus, pharynx, larynx and breast [26]. The effect of overdrinking also affects human behavior, significantly increasing the risk after drinking. Alcohol is the third leading preventable cause of death in the United States and nearly 88000 deaths can be directly or indirectly linked to alcohol consumption every year [27]. Moreover, from an economic point of view, the burden of solving Alcohol misuse problems in the United States cost $233.5 billion in 2006, three-quarters of which was related to binge drinking.

Thus, avoiding overdrinking saves life, time and money. However, even though most people are aware of the harm of excess alcohol consumption, it is difficult to prevent overdrinking. The situation is further compounded by the fact that alcohol can affect the human brain, leading people to make wrong decisions. And things typically get worse as people drink more.

Individuals who monitor their alcohol consumption generally avoid overdrinking. However, manual recording presents a significant burden. Thus, novel, autonomous methods of detecting alcohol consumption are needed. The standard quantified unit of measuring alcohol consumption is Blood Alcohol Concentration (BAC) or Breath Alcohol Concentration (BrAC), which is the amount of alcohol in the bloodstream or in the breath, respectively. BAC is expressed as the weight of ethanol, measured in grams, in 100 milliliters of blood. BrAC is the weight of ethanol in 210 liters of breath [21]. When a person drinks alcohol, it can either spread into the blood, or be released through their breath [22]. BAC and BrAC can be measured by breath, blood, or urine tests [21].

[1] NIAAA defines binge drinking as a pattern of drinking that brings blood alcohol concentration (BAC) levels to 0.08 g/dL. This typically occurs after 4 drinks for women and 5 drinks for men—in about 2 hours [28].

1.2 Motivation for a Better Way to Detect Alcohol Consumption levels

Typical ways of detecting a person's alcohol consumption status are his/her BAC or BrAC measurement. BAC and BrAC are direct and accurate measures of alcohol consumption, but requiring extra devices such as breathalyzers and manual operation. The user has to carry such devices with them and test their BAC or BrAC levels after each drink. This is obviously annoying, burdensome and easy to forget. Due to their inconvenience and required participation from the user, BAC devices such as the breathalyzer are limited in how much they can reduce the number of overdrinking cases, or problems. Sensors on wearable devices have also been proposed but users may forget to wear such sensors.

Researchers have also investigated other indirect methods of detecting alcohol consumption including gait, which is defined as a coordinated effort by the brain and other muscles to produce mobility in an effort to go somewhere [9]. Figure 1 shows the human gait cycle. Alcohol intoxication has a significant effect on how the human body operates. There are two major organs in the human body which respond sensitively to alcohol consumption: the heart and the brain [20]. Approximately ten minutes after the initial alcohol consumption, the heart rate begins to increase in order to filter out the toxins from the bloodstream through the kidneys. After about twenty minutes, the alcohol is able to penetrate the blood-brain barrier causing noticeable impacts to cognitive and neuromotor functions. And human gait is among these functions. Alcohol impairment significantly impacts this coordination and can dramatically impact the ability to walk, jog, or run, and finally reflects in human gait. Gait analysis has previously been found to be useful in the detection of many diseases and impairment, leading to attempts to employ it in the field of alcohol consumption detection.

Sensors, such as accelerometers and gyroscopes that are now integrated into many smartphones, have been widely used to assess gait for many years [56] [57]. Due to their improved measurement accuracy, ease, and affordability, and reliability accelerometers have become the most popular gait measurement sensors. Accelerometers have been found to be reliable in gait analysis, and is robust over several days [58], with changes in walking speed [59] and surfaces [60]. Besides, the accelerometer is also the most popular sensor in smartphones.

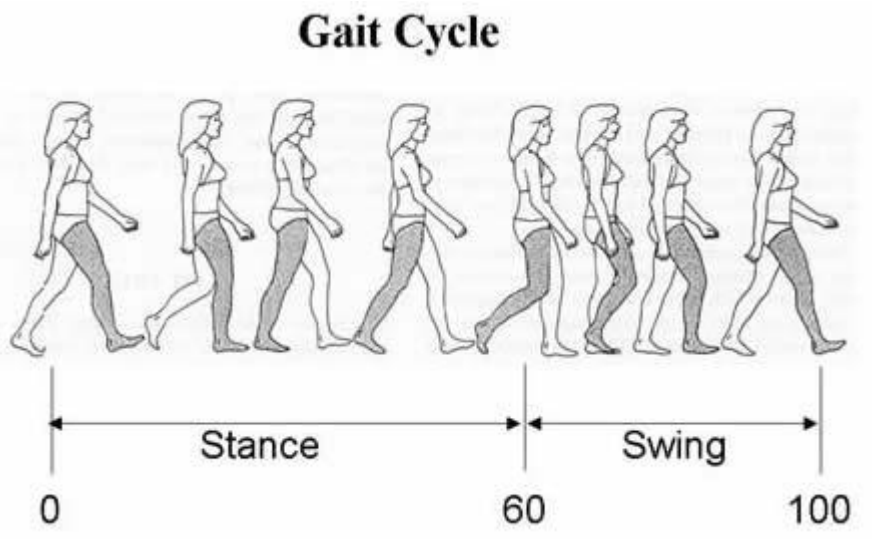

Figure 1 The Human Gait Cycle [43]

Figure 2 People Walking with Smartphone in Pocket (Back)

Arnold *et al* [1] in our research lab demonstrated the feasibility of detection of the number of drinks consumed by a user based on smartphone accelerometer data. This idea is inspired by the wide use of smartphones in our daily life. People usually walk with a smartphone in their pocket (see Figure 2). So we can use a smartphone app that reads smartphone sensor data (e.g. accelerometer and gyroscope) to analyze gait, instead of a separate annoying device. Arnold *et al* [1] created the AlcoGait app that could autonomously collect data by itself and recognize whether the user is sober, binge

drinking or heavily drunk. The user was then notified to alert them of their drinking status and avoid overdrinking or alcohol-related problems. The alcogait app extracted 11 time and frequency domain features from accelerometer data gathered from the user's smartphone, and classified them using a supervised learning machine learning approach, achieving an accuracy of only 57%.

1.3 Goals of This Thesis

This thesis builds on the work of Arnold et al. Specifically, their work considered only 11 features in the time and frequency domain. However, several promising features have also been proposed to analyze gait in the statistical, information-theoretic and wavelet domains. Inspired by the comparison on features by Sedjic et al [5], a comprehensive list of signal processing features is evaluated for how much they improve the classification accuracy of alcohol consumption from accelerometer gait data. While gyroscopes have also been explored for gait analysis, many smartphone currently do not have gyroscopes. However, accelerometers, the most basic gravity sensors have been integrated into many smartphones. Thus, in this thesis, we analyze, classify and compare only accelerometer data for their performance in detecting alcohol consumption from gait data.

Since relatively few published papers focused on alcohol detection in the area of gait analysis, we have expanded our list of evaluated features by including features used for smartphone detection of similar conditions based on gait. Table 1 lists all 28 features we considered along with health conditions in which they were originally used to detect. These conditions included Parkinson's disease, peripheral neuropathy, coronary artery, and general neural and heart health problems. These ailments all alter gait in some fashion that could be considered somewhat similar to the alcohol consumption problem. As stated in section 1.2, alcohol consumption affects the human heart and brain, and eventually alters subjects' gait. Thus it is reasonable to investigate these features for the task of alcohol consumption detection.

Table 1 Accelerometer Gait Features and Their Original Use Cases

	Featrue Name	Applied Cases
1	Number of Steps	Alcohol Usage [2], Parkinson's Disease [51] [54]
2	Average Step Time	Alcohol Usage [2] [10], Parkinson's Disease [51] [54]
3	Average Cadence	Alcohol Usage [2], Parkinson's Disease [51]

4	Skewness	Alcohol Usage [2], absolute activation of paraspinal muscles assessment [52]
5	Kurtosis	Alcohol Usage [2], Neuron Discharge [53]
6	Coefficient of Variation of Step Time	Parkinson's Disease [5], peripheral neuropathy [5]
7	Harmonic Ratio	Parkinson's Disease [5], peripheral neuropathy [5]
8	Average Step Length	Alcohol Usage [2], Parkinson's Disease [54]
9	Gait Velocity	Alcohol Usage [2], Parkinson's Disease [54]
10	Minimum and Maximum Difference	Parkinson's Disease [4]
11	Standard Deviation	Parkinson's Disease [4] [5], peripheral neuropathy [5], absolute activation of paraspinal muscles assessment [52]
12	Root Mean Square	Parkinson's Disease [4]
13	Entropy Rate	Parkinson's Disease [4] [5], peripheral neuropathy [5], neural control [13], Heart [14]
14	Regression Line for Local Maxima and Minima	Parkinson's Disease [4]
15	Average Power	Alcohol Usage [2], absolute activation of paraspinal muscles assessment [52]
16	Ratio of Spectral Peak	Alcohol Usage [2]
17	Signal Noise Ratio	Alcohol Usage [2], coronary artery [55]
18	Total Harmonic Distortion	Alcohol Usage [2]
19	Energy in Band 0.5 to 3 Hz	Parkinson's Disease [4]
20	Windowed Energy in Band 0.5 to 3 Hz	Parkinson's Disease [4]
21	Peak Frequency	Parkinson's Disease [5], peripheral neuropathy [5], absolute activation of paraspinal muscles assessment [52]
22	Spectral Centroid	Parkinson's Disease [5], peripheral neuropathy [5]
23	Bandwidth	Parkinson's Disease [5], peripheral neuropathy [5]
24	Regression Line for windowed Energy	Parkinson's Disease [4]

25	Wavelet Bandwidth	Parkinson's Disease [5], peripheral neuropathy [5]
26	Wavelet Entropy Rate	Parkinson's Disease [5] [15], dysphagia [15], neural control/condition [15] [16]
27	Zeroth-Lag Cross-Correlation Coefficient	Parkinson's Disease [5], peripheral neuropathy [5]
28	Lampel-Ziv Complexity	Parkinson's Disease [5], peripheral neuropathy [5], electroencephalogram [19]

Additionally, we believed that feature performance may also be affected by the methods used in generating them. As such, we explored 3 alternate approaches (welch power spectral density, FFT and DCT) for generating the Ratio of Spectral Peak feature. Including these alternate implementation methods brought our total number of compared features to a total of 30 features.

In this thesis, 27 signal processing features from the smart phone accelerometer data are compared first using Correlation-Based Feature Selection (CFS) [34] wherein each feature's correlation with alcohol consumption level and p-value are computed. The features that are most strongly correlated with BAC levels (p-value < 0.05) have the highest predictive value and are then used for classification of alcohol consumption levels. The correlation values of all 27 features were ranked and analyzed individually as well as in families of signal processing features. To account for the fact that different people have different walk patterns even before drinking, all features are also normalized by dividing each subject's observed feature by its value in their sober walk.

Finally, the effect of each feature type on classification accuracy is evaluated individually as well as in families of signal processing features. The accuracy of different types of machine learning classifiers such as Random Forest, SVM and Naïve Bayes are compared for different families of features. Detailed results of classification are presented including accuracy, precision, recall, ROC curves and confusion matrices.

This work will contribute to the area of alcohol consumption detection from anomalies in human gait and will help future investigators and industry to select the best features for alcohol consumption detection.

1.4 Thesis Organization

This thesis is organized as follows: Chapter 2 gives some background including definitions of the signal processing features investigated. Chapter 3 explains our data gathering methodology, the input gait data set as well as post-processing methods and methodology. Chapter 4 shows the result of the post process and correlation between features and alcohol consumption levels, and the results of machine learning classification. Chapter 5 makes conclusions and reasoning based on the results from chapter 4.

2. Background and Related Work

This chapter describes related work and further clarifies the contributions of this thesis. Definitions of all signal processing features extracted and tested in this thesis are then presented. These features are grouped by their categories. Some features may have multiple categories, which are noted in footnotes.

2.1 Related Work

Related work in three related areas is now reviewed: alcohol detection devices, alcohol calculation app, and other gait-based analysis researches.

Alcohol detection devices: There are several alcohol detection device products in the market, such as SCRAM and Kisai Intoxicated LCD Watch (shown in the Figure 3 below). The former, SCRAM Continuous Alcohol Monitoring [29] is a commercial device that is worn continuously around the ankle. It measures the users' BAC levels by sampling their perspiration every 30 minutes. The latter, Kisai Intoxicated LCD Watch is a breathalyzer watch developed by TokyoFlash Japan [30]. It is a normal watch plus a built-in breathalyzer on its side, which determines BrAC level when the user breathes into it. Their disadvantages include being an additional device and having complex operation, respectively. Our gait detection approach works passively without extra burden on the user, only requiring a smart phone, which is always carried by most people, and does not need any further effort beyond a training session.

Figure 3 Kisai Intoxicated LCD Watch [30]

Alcohol Calculation apps: There are several apps in the mobile app market, but many require manual user input in order to calculate or estimate the user's alcohol consumption level, which is burdensome. Apps, such as IntelliDrink [31] (shown in the figure 4 below) and AlcoDroid Alcohol Tracker [32], returns the user's BAC level based on the number of drinks, time elapsed and the user's personal profile. But a major problem is that they at least require the user to put in their number of drinks before complete their estimation. However, people may forget to put in this information while drinking, for reasons of pleasure or depression. Another app, proposed by Kao et al, also used smartphone based sensors to classify alcohol consumption of users, and is similar to ours. However, it can only determine whether the user has consumed alcohol (Yes/No), but not how heavily the user drinks (number of drinks or BAC levels). Our gait classification approach will run passively on the user's smartphone, presenting minimal burden and will detect BAC levels of users.

Figure 4 IntelliDrink - a sophisticated blood alcohol content (BAC) calculator [32]

Researchers have also extracted health-related information from gait for disease-oriented applications [5]. Klucken et al applied gait analysis to characterize the movement of patients afflicted with Parkinson's disease [4]. However, few researchers are focusing on gait classification for the purposes of alcohol consumption detection, which is the topic of my research.

Finally compared to previous work in our research lab by Arnold *et al*, my work introduces a more comprehensive list of 27 signal processing features (compared to 11 by them) that are extracted from accelerometer gait signals, which are classified and analyzed for their potential in detecting alcohol

consumption levels. Where applicable, alternate extraction methods (e.g. FFT vs DCT) are evaluated in this thesis.

2.2 Definitions of Signal Processing Features

The signal processing features investigated are used to extract features from accelerometer data. A sample of the accelerometer data is shown in figure 5. To illustrate feature calculation, the magnitude of the acceleration vector is also calculated and shown. This magnitude will be used as a variable x in calculating our time domain features.

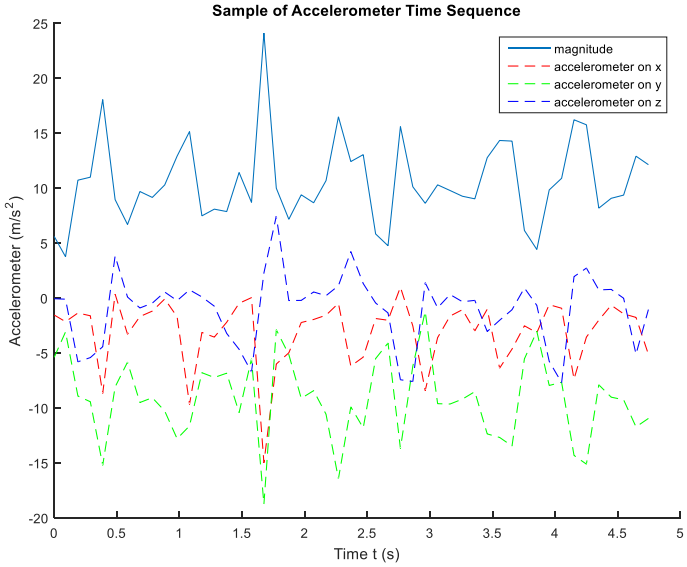

Figure 5 Sample Figure of Accelerometer Time Sequence, x, y and z acceleration are in red, green and blue dash. The solid line is the magnitude of acceleration, which is equal to $\sqrt{x^2 + y^2 + z^2}$

2.2.1 Time Domain Features

The following table summarizes the time domain features investigated in this thesis.

Table 2 Time Domain Features

Feature	Abbr. of Feature	Description
Number of Steps	numSteps	The number of steps taken in a given time interval [2] [9]
Average Step Time	averageStepTime	The average time elapsed for each step [2] [10]
Average Cadence	averageCadence	The ratio of the total number of steps by the total time

		[2] [9]
Skewness	skewness	Asymmetry of the signal distribution [2] [5] [9]
Kurtosis	kurtosis	The extent to which the distribution of signal amplitudes lies predominantly on the left of the mean amplitude [2] [5] [9]
Coefficient of Variation of Step Time	coef of var of stepTime	Within-subject standard deviation of the stride interval divided by the mean stride interval [5] [11]
Harmonic Ratio	harmonic ratio	Harmonic Ratio quantifies the harmonic composition of the accelerations for a given stride via DFT [5] [12]
Average Step Length	averageStepLength	The average distance covered by each step [2] [10]
Gait Velocity	gaitVelocity	The ratio of the total distance covered by the total time [2] [9]
Minimum and Maximum Difference	minMaxDiff	Global maximum of one step minus global minimum of one step, averaged over all steps of one subject [4]
Standard Deviation	std	Measure for signal spreading, defined as the square of standard deviation [4] [5]
Root Mean Square	rms	Root Mean Square or quadratic mean is a statistical measure [4]
Entropy Rate	entropy rate	the uncertainty measure of the signal, and the regularity of a signal when anticipated that consecutive data points are related [4] [5] [13] [14]
Regression Line for Local Maxima and Minima	-	egression line of all local minima and maxima in the signal sequence [4]

2.2.1.1 Number of Steps (numSteps)

This feature is defined as the number of steps taken in a given time interval. We generate it using the step detection method – Local Peaks [2] [9]. The number of steps in a time domain sequence is equal to the number of those local maxima which is above the average value of the entire sequence. This is a feature carried forward from Arnold *et al's* work [2]. Figure 6 shows time series accelerometer data and how we detect the 3 steps in it.

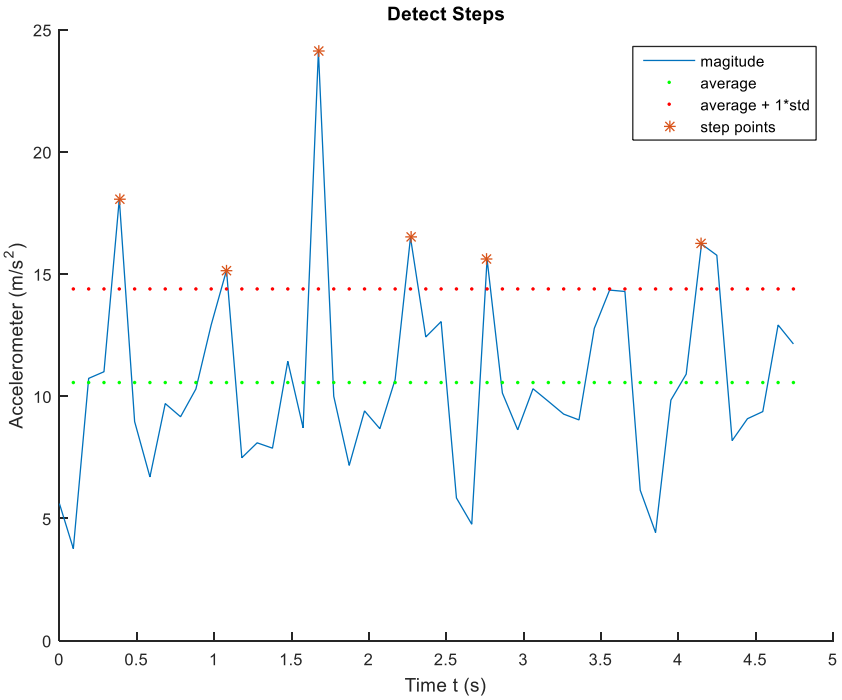

Figure 6 Example of Step Detection, via finding local peaks above average plus one standard deviation. The stars stand for points of detected steps.

2.2.1.2 Average Step Time (averageStepTime)

This feature is defined as the average time elapsed for each step [2] [10]. This is a feature carried forward from Arnold et al's work [2].

$$averageStepTime = \frac{time}{\#steps}$$

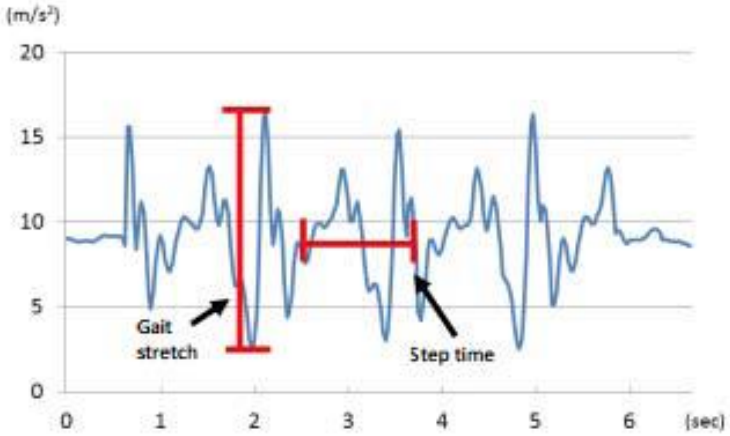

Figure 7 Example Data showing Gait Stretch and Step Time [10]

2.2.1.3 Average Cadence (averageCadence)

This feature is defined as the Ratio of the total number of steps and the total time taken [2] [9]. This is a feature carried forward from Arnold et al's work [2].

$$averageCadence = \frac{\#steps}{time}$$

2.2.1.4 Skewness (skewness)[2]

This feature is defined as the asymmetry of the signal distribution [2] [5] [9]. If the value of skewness is negative, distribution of signal amplitudes lies predominantly on the right of the mean amplitude. And if it is positive, the distribution of signal amplitudes lies predominantly on the left. This is a feature carried forward from Arnold et al's work [2].

$$skewness = \frac{\frac{1}{n}\sum(x_i - \mu_x)^3}{\left[\frac{1}{n}\sum(x_i - \mu_x)^2\right]^{3/2}}$$

x_i refers to a data sequence in which skewness is to be calculated, and it refers to the accelerometer data in this thesis. μ_x refers to the average of all x_i.

2.2.1.5 Kurtosis (kurtosis)[3]

This feature is defined as the extent to which the distribution of signal amplitudes lies predominantly on the left of the mean amplitude [2] [5] [9]. A higher kurtosis values indicates the distribution is more peaked, with infrequent, extreme deviations. This feature was initially explored in the work of Arnold et al [2].

$$kurtosis = \frac{\frac{1}{n}\sum(x_i - \mu_x)^4}{\left[\frac{1}{n}\sum(x_i - \mu_x)^2\right]^2}$$

x_i refers to a data sequence in which kurtosis is to be calculated, and it refers to the accelerometer data in this thesis. μ_x refers to the average of all x_i.

2.2.1.6 Coefficient of Variation of Step Time (coef of var of stepTime)

This feature is defined as the within-subject standard deviation of the stride interval divided by the mean stride interval [5] [11]. The stride interval is the time between two steps (see Figure 7). It is

[2] Also mentioned as Statistical Features in [5]

[3] Also mentioned as Statistical Features in [5]

usually presented in a percentage format. This feature showed promising results in [5]. The coefficient of variation represents the variance of step times during a target walking sequence, and will possibly also capture alcohol-induced gait anomalies in this thesis.

$$coef\ of\ var\ of\ stepTime = \frac{\sqrt{\frac{1}{n}\sum(interval_i - \mu_{interval})^2}}{\mu_{interval}}$$

$interval_i$ refers to a sequence of stride intervals. $\mu_{interval}$ refers to the average of all $interval_i$.

2.2.1.7 Harmonic Ratio (harmonic ratio)

This feature is defined as Harmonic Ratio (HR) quantifies the harmonic composition of the accelerations for a given stride via DFT [5] [12]. HRs are calculated using the first 20 harmonic coefficients; higher values are interpreted as greater walking smoothness. Harmonics represents the composition of the target signal. This feature showed promising results in [5], leading us to want to explore it further.

$$harmonic\ radio = \frac{\sum_{i=1,3,5,\ldots} V_i}{\sum_{j=2,4,6,\ldots} V_j}$$

V_i refers to the amplitude of odd ordered harmonic frequency in the frequency domain, while V_j refers to that of even-ordered harmonic frequency.

2.2.1.8 Average Step Length (averageStepLength)

This feature is defined as the average distance covered by each step [2] [10].

$$averageStepLength = \frac{distance}{\#steps}$$

However, it is hard to calculate a direct value of distance from only accelerometer sensor data. Thus this feature was generated in an indirect way. Since there is a linear relationship between step frequency and step length [7], this feature was generated by using the value of step time of this sequence.

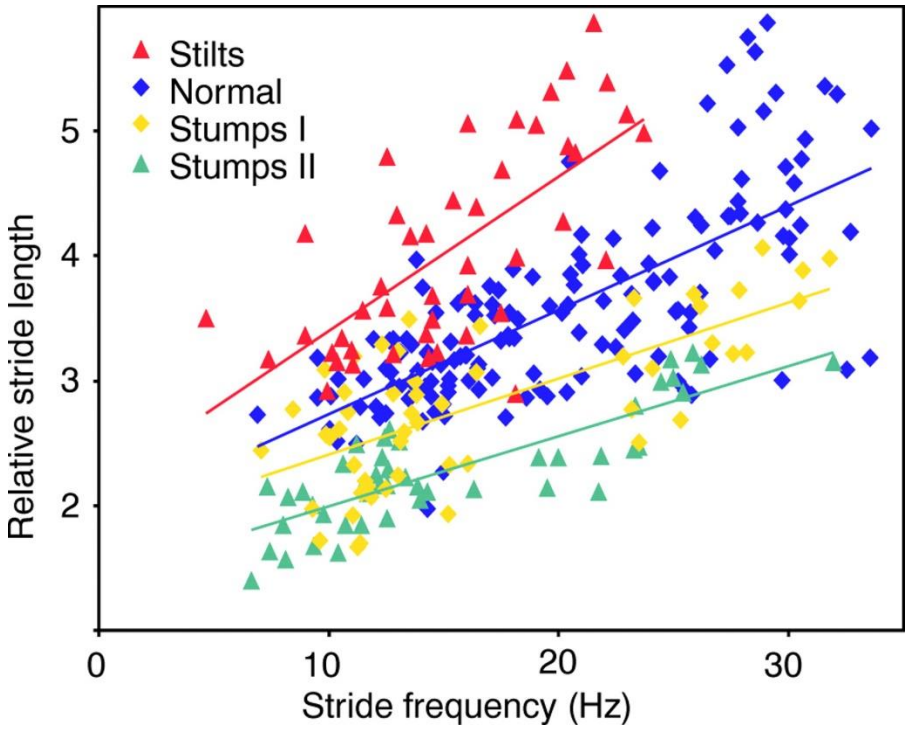

Figure 8 Stride Frequency vs Stride Length relationship from [8]

And from Wittlinger et al [8], this feature is defined in this thesis as:

$$averageStepLength = \frac{0.084}{averageStepTime} + 1.89$$

The constants chosen (0.084 and 1.89) above are according to the Wittlinger et al's conclusion [8]. The gradient and intercept may change among individuals, however, a simple local training can help set gradient and intercept, eliminating the effect of individuals.

2.2.1.9 Gait Velocity (gaitVelocity)

This feature is defined as the ratio of the total distance covered divided by the total time [2] [9].

$$gaitVelocity = \frac{distance}{time}$$

However, it is hard to get a direct value of distance from only accelerometer sensor data. This feature was generated in an indirect way. Since there is a linear relationship between step time and step length [7], this feature was generated using the value of step time of this sequence by Wittlinger et al [8]. This feature is defined in this thesis as:

$$gaitVelocity = \frac{\left(\frac{0.084}{averageStepTime} + 1.89\right)}{averageStepTime}$$

The gradient and intercept may change among individuals, however, a simple local training can be used to set gradient and intercept, getting rid of the effect of individuals.

2.2.1.10 Minimum and Maximum Difference (minMaxDiff)[4]

This feature is defined as: Global maximum of one step minus the global minimum of one step, averaged over all steps of one subject [4]. It was considered in this thesis because of its promising performance in [4] although it was used there for gyroscope rather than accelerometer data in [4].

$$minMaxDiff = \max(x_i) - \min(x_i)$$

2.2.1.11 Standard Deviation (std)[5]

This feature is defined as: Measure for signal spreading, defined as the square of standard deviation [4] [5]. Higher values indicate a greater spread of amplitude values. This feature is a typical feature for all statistical data, and it performed well in [4], leading us to include it in our list of signal processing features to be evaluated in this thesis.

$$std = \sqrt{\frac{1}{n}\sum(x_i - \mu_x)^2}$$

x_i refers to a data sequence in which we want to calculate standard deviation, and it refers to the accelerometer data in this thesis. μ_x refers to the average of all x_i.

2.2.1.12 Root Mean Square (rms)[6]

This feature is defined as the Root Mean Square or quadratic mean, which is a statistical measure [4]. This feature is a typical feature for all statistical data, and it performed well in [4], leading us to include it the list of features to be evaluated in this thesis.

$$rms = \sqrt{\frac{1}{n}\sum {x_i}^2}$$

x_i refers to a data sequence in which we want to calculate RMS, and it refers to the accelerometer data here.

---

[4] Mentioned as Step Dependent Features in [4]

[5] Mentioned as Sequence Dependent Features in [4], and as Statistical Features in [5]

[6] Mentioned as Sequence Dependent Features in [4]

2.2.1.13 Entropy Rate (entropy rate)[7]

This feature is chosen because it is considered as the uncertainty measure of the signal, and the regularity of a signal when it is anticipated that consecutive data points are related [4] [5] [13] [14]. Its values range from 0 to 1 where 0 refers to maximum randomness/no relationship among consecutive data points, and 1 refers to maximum regularity.

$$entropy\ rate = -\sum possibility_{unique\ freq} \times log_2(possibility_{unique\ freq})$$

2.2.1.14 Regression Line for Local Maxima and Minima[8]

This feature is defined as the Regression line of all local minima and maxima in the signal sequence [4]. It is considered here because of its promising performance in [4] although in that work it was used on gyroscope rather than accelerometer data.

2.2.2 Frequency Domain Features

The following table lists all frequency domain features investigated in this thesis.

Table 3 Frequency Domain Features

Feature	Abbr. of Feature	Description
Average Power	average power	the mean of the total power underneath the curve of the PSD estimate for a signal [2] [9]
Ratio of Spectral Peak (with 3 derivatives: Welch, FFT and DCT)	ratioSpectralPeak	Ratio of the energies of low and high frequency bands [2] [9]
Signal Noise Ratio	snr	Power of whole signal over power of its computed noise [2]
Total Harmonic Distortion	thd	Distortion of the whole signal compared to its harmonics [2]
Energy in Band 0.5 to 3Hz	energy in _5 to 3	Energy in a frequency band describes parts of distinct frequencies in the signal, and the frequency range is recommended as 0.5Hz to 3Hz [4]
Windowed Energy in Band 0.5 to 3Hz	windowed energy in _5 to 3	Energy in frequency band of 5 second windows with an overlap of 2.5 seconds, windows from complete signal sequence are averaged [4]
Peak Frequency	peakFreq	The maximum spectral power [5]
Spectral Centroid	spectralCentroid	The frequency that divides the spectral power distribution into two equal parts [5]

[7] Mentioned as Sequence Dependent Features in [4], and as Information-Theoretic Features in [5]

[8] Mentioned as Sequence Dependent Features in [4]

Bandwidth	bandwidth	The difference between the uppermost and lower most frequencies/range of frequencies in the signal [5]
Regression Line for windowed Energy	-	Regression line of energy values from window (2.5 s) moved through signal sequence [4]

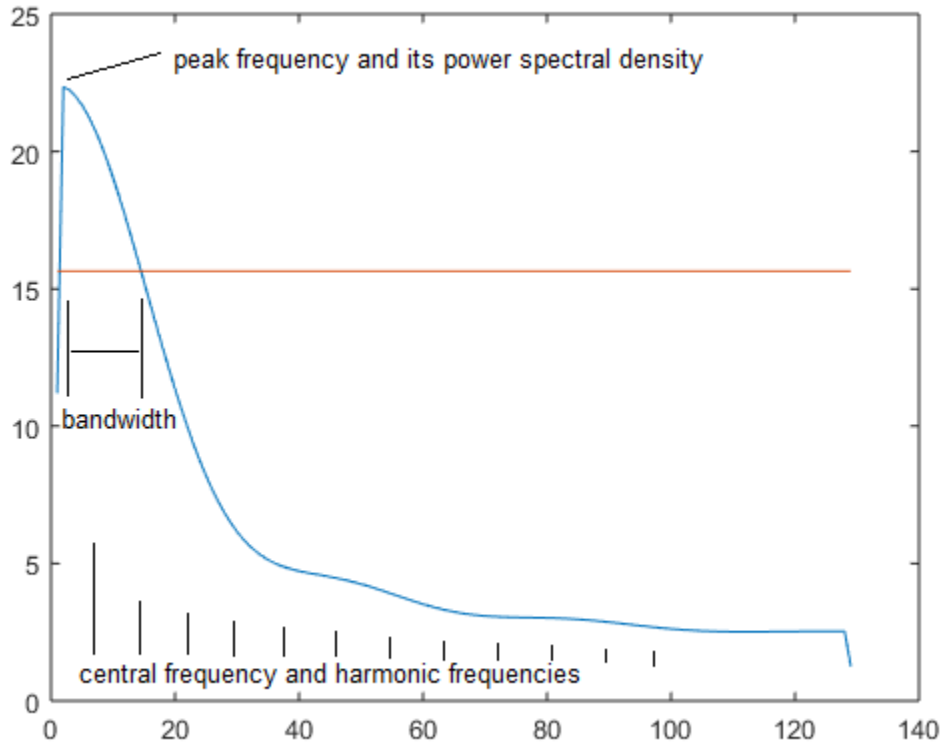

Figure 9 Frequency Domain Power Spectral Density and Its Concepts

2.2.2.1 Average Power (average power)

This feature is defined as the mean of the total power underneath the curve of the PSD estimate for a signal [2] [9].

$$average\ power = \frac{total\ power\ of\ the\ signal}{bandwidth\ of\ the\ signal}$$

2.2.2.2 Ratio of Spectral Peak

This feature is defined as the ratio of the energies of low and high frequency bands [2] [9].

$$ratioSpectralPeak = \frac{\max(power_{freq})}{mean(power_{freq})}$$

Different DFT methods were attempted to generate this feature, in order to find which of the DFT methods could improve the performance of frequency domain features. In this thesis, the effect of

different time-frequency transform methods on a feature are also investigated. In total, 3 alternative methods (default Welch transform, using FFT, and using DCT) were implemented to calculate this feature. Of the 3 approaches, I believe FFT will perform the best for this thesis. Welch, which refers to the Welch's overlapped segment averaging estimator, is usually a good approach in many conditions. But due to the segmentation in preprocessing and the limit of data length, FFT would do better. To prove it, I tested all 3 approaches, and the result can be found in Section 4.2.2.

2.2.2.3 Signal Noise Ratio (SNR)

This feature is defined as the power of the whole signal divided by the power of its computed noise [2].

$$snr = \frac{power_{signal}}{power_{noise}}$$

2.2.2.4 Total Harmonic Distortion (thd)

This feature is defined as the distortion of the whole signal compared to its harmonics [2]. The harmonics are illustrated in Figure 9 above.

$$thd = \frac{\sqrt{\sum_{i=2,3,4,5} V_i^2}}{V_1}$$

V_i refers to the ith harmonic frequency in the frequency domain, while V_1 is the base frequency.

2.2.2.5 Energy in Band 0.5 to 3 Hz (energy in _5 to 3)

This feature is defined as the energy in a frequency band and describes parts of distinct frequencies in the signal. The frequency range is recommended as 0.5Hz to 3Hz [4]. Typical frequency bands for specific movements can be defined. It is considered here because of its promising performance in [4] although in that work, it was applied to gyroscope rather than accelerometer data.

$$energy\ in\ 0.5\ to\ 3 = \int_{0.5}^{3} psd_f\ df$$

psd_f refers to the power spectral density of frequency. And frequency range is from 0.5 Hz to 3 Hz. In discrete signal processing as in the accelerometer data analyzed in this thesis, the integral is converted into sum.

2.2.2.6 Windowed Energy in Band 0.5 to 3 Hz (windowed energy in _5 to 3)

This feature is defined as the energy in a frequency band of 5 second windows with an overlap of 2.5 seconds, windows from complete signal sequences are averaged [4]. It is considered in this thesis

because of its promising performance in [4] although in that work, it was applied to gyroscope rather than accelerometer data.

$$windowed\ energy\ in\ 0.5\ to\ 3 = \int_{0.5}^{3} windowed\ psd_f\ df$$

$windedewed\ psd_f$ refers to the windowed power spectral density of frequency. And frequency range is from 0.5 Hz to 3 Hz. In discrete signal processing as in this thesis, the integral is converted into sum.

2.2.2.7 Peak Frequency (peakFreq)

This feature is defined as the maximum spectral power [5]. It is chosen because it denotes the frequency at which the maximum spectral power occurred and worked promisingly in [5]. Although it is named as peak frequency in [5], this feature is in fact the power of peak frequency. See Figure 9 above.

$$peakFreq = \max(power_f)$$

2.2.2.8 Spectral Centroid (spectralCentroid)

This feature is defined as the frequency that divides the spectral power distribution into two equal parts [5]. This feature is similar to peak Frequency but is another way to explore power distribution.

$$spectralCentroid = \frac{\sum f \times {power_f}^2}{\sum {power_f}^2}$$

2.2.2.9 Bandwidth (bandwidth)

This feature is defined as the difference between the uppermost and lower most frequencies/range of frequencies in the signal [5]. This is a typical feature for all frequency domain analysis. See Figure 9 above.

$$bandwidth = \frac{\sum (f - spectralCentroid)^2 \times {power_f}^2}{\sum {power_f}^2}$$

2.2.2.10 Regression Line for windowed Energy

This feature is defined as the regression line of energy values from 2.5 second windows moved through the signal sequence [4]. It is considered here because of its promising performance in [4] although in that work it was applied to gyroscope rather than accelerometer data.

2.2.3 Wavelet Domain Features

Wavelet domain features illustrates the property of the signal in time-frequency domain. They

are generated from wavelet transform. The Following Figure shows an example of continuous Cauchy Wavelet Transform [61].

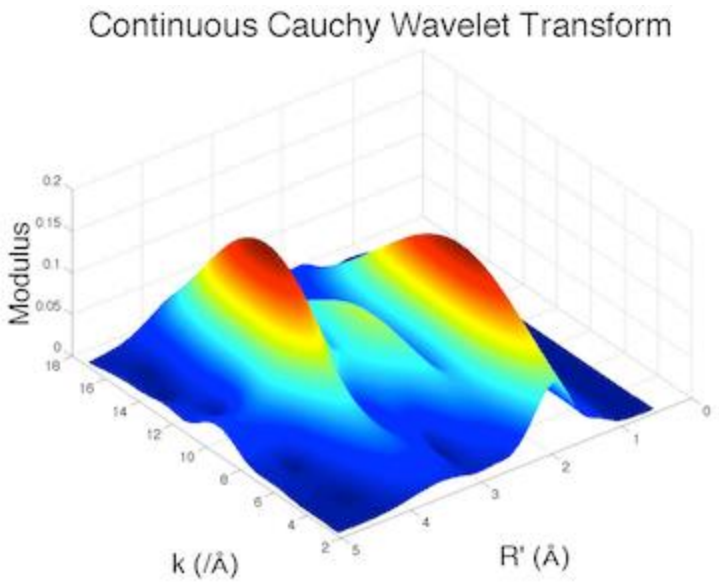

Figure 10 an example of Continuous Cauchy Wavelet Transform

The following table lists all wavelet domain features investigated in this thesis.

Table 4 Wavelet Domain Features

Feature	Abbr. of Feature	Description
Wavelet Bandwidth	wavelet band	The relative energy contribution in a time-frequency band [5]
Wavelet Entropy Rate	wavelet entropy	Wavelet entropy represents signal disorder in the time-frequency domain [5] [15] [16]

2.2.3.1 Wavelet Bandwidth (wavelet band)

This feature is defined as the relative energy contribution in a time-frequency band [5]. This is a repeating feature in the wavelet domain.

$$[cA, cD] = dwt(x, 'db1');$$

$$wavelet\ band = cA' * cA/(cA' * cA + cD' * cD);$$

dwt refers to discrete wavelet transform. It computes the approximation coefficients vector cA and detail coefficients vector cD. Using these vectors, which describe the wavelet property of the sequence, we can calculated the wavelet bandwidth [33].

2.2.3.2 Wavelet Entropy Rate (wavelet entropy)

This feature is defined as the wavelet entropy and represents signal disorder in the time-frequency domain [5] [15] [16]. High values represent disordered behavior with significant equivalent contributions from all frequency bands. This is a repeating feature in the wavelet domain.

$$entropy\ rate = -\sum possibility_{unique\ freq} \times log_2(possibility_{unique\ freq})$$

2.2.4 Statistical Features

The following table lists all statistical features investigated in this thesis.

Table 5 Statistical Features

Feature	Abbr. of Feature	Description
Zeroth-Lag Cross-Correlation Coefficient	cross correlation	The agreement or similarity between 2 directional acceleration signals [5]
Kurtosis	kurtosis	The extent to which the distribution of signal amplitudes lies predominantly on the left of the mean amplitude [2] [5] [9]
Standard Deviation	std	Measure for signal spreading, defined as the square of standard deviation [4] [5]

2.2.4.1 Zeroth-Lag Cross-Correlation Coefficient (cross correlation)

This feature is defined as the agreement or similarity between 2 directional acceleration signals [5]. Its value ranges from 0 to 1 where 0 indicates no similarity and 1 indicates identical signals. It was chosen because it worked well in [5] leading us to consider it in this thesis.

$$cross\ correlation = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum(x_i - \mu_x)^2 \sum(y_i - \mu_y)^2}}$$

x_i refers to a sequence of data in which cross correlation is to be calculated, and it refers to the accelerometer data here. μ_x refers to the average of all x_i.

The following figure shows an example of general cross-correlation [62].

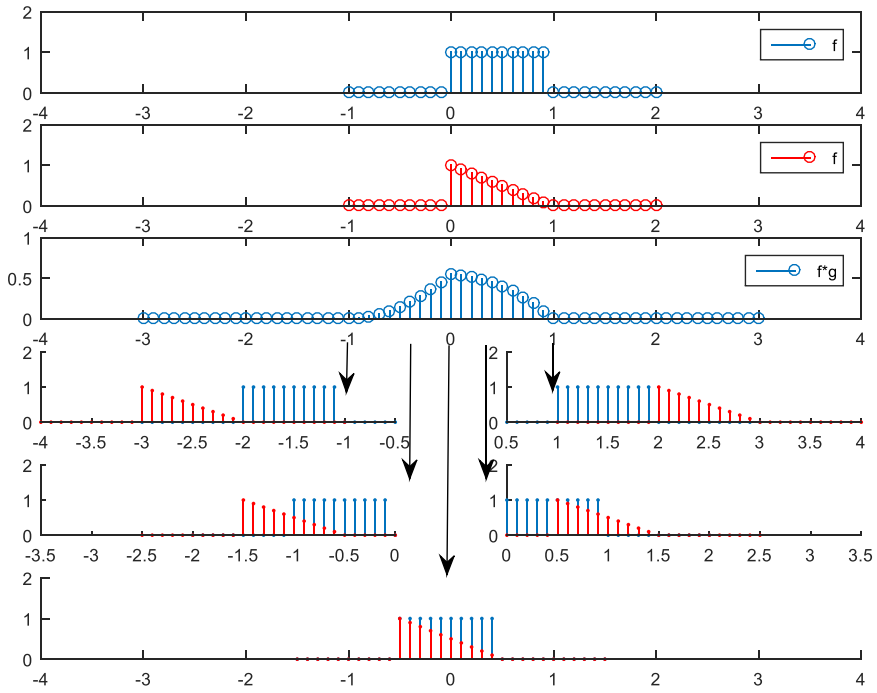

Figure 11 Example of Cross-Correlation

2.2.5 Information-Theoretic Features

The following table lists all information-theoretic features investigated in this thesis.

Table 6 Information-Theoretic Features

Feature	Abbr. of Feature	Description
Lampel-Ziv Complexit	-	The complexity-predictability of the signal [5] [17] [18] [19]
Entropy Rate	entropy rate	the uncertainty measure of the signal, and the regularity of a signal when anticipated that consecutive data points are related [4] [5] [13] [14]

2.2.5.1 Lampel-Ziv Complexity

This feature is defined as the complexity-predictability of the signal [5] [17] [18] [19].

2.2.6 Feature Summary

A total of 30 features were introduced in this thesis, while 3 of them were not test (see Future Work). And 11 among them were carried forward from Arnold et al [2]. The table below list new features investigated in bold, as well as features carried-forward in bold and Italics in the bottom half of the table.

Table 7 Table of all Features (new in Bold, and carried-forward in *Bold and Italics*)

Feature	Description
Coefficient of Variation of Step Time	Within-subject standard deviation of the stride interval divided by the mean stride interval [5] [11]
Harmonic Ratio	Harmonic Ratio quantifies the harmonic composition of the accelerations for a given stride via DFT [5] [12]
Minimum and Maximum Difference	Global maximum of one step minus global minimum of one step, averaged over all steps of one subject [4]
Standard Deviation	Measure for signal spreading, defined as the square of standard deviation [4] [5]
Root Mean Square	Root Mean Square or quadratic mean is a statistical measure [4]
Entropy Rate	the uncertainty measure of the signal, and the regularity of a signal when anticipated that consecutive data points are related [4] [5] [13] [14]
Regression Line for Local Maxima and Minima	egression line of all local minima and maxima in the signal sequence [4]
Ratio of Spectral Peak (with 2 new derivatives: FFT and DCT)	Ratio of the energies of low and high frequency bands [2] [9]
Energy in Band 0.5 to 3Hz	Energy in a frequency band describes parts of distinct frequencies in the signal, and the frequency range is recommended as 0.5Hz to 3Hz [4]
Windowed Energy in Band 0.5 to 3Hz	Energy in frequency band of 5 second windows with an overlap of 2.5 seconds, windows from complete signal sequence are averaged [4]
Peak Frequency	The maximum spectral power [5]
Spectral Centroid	The frequency that divides the spectral power distribution into two equal parts [5]
Bandwidth	The difference between the uppermost and lower most frequencies/range of frequencies in the signal [5]
Regression Line for windowed Energy	Regression line of energy values from window (2.5 s) moved through signal sequence [4]
Wavelet Bandwidth	The relative energy contribution in a time-frequency band [5]
Wavelet Entropy Rate	Wavelet entropy represents signal disorder in the time-frequency domain [5] [15] [16]
Zeroth-Lag Cross-Correlation Coefficient	The agreement or similarity between 2 directional acceleration signals [5]
Lampel-Ziv Complexit	The complexity-predictability of the signal [5] [17] [18] [19]
Number of Steps	*The number of steps taken in a given time interval [2] [9]*
Average Step Time	*The average time elapsed for each step [2] [10]*
Average Cadence	*The ratio of the total number of steps by the total time [2] [9]*
Average Step Length	*The average distance covered by each step [2] [10]*
Gait Velocity	*The ratio of the total distance covered by the total time [2] [9]*
Skewness	*Asymmetry of the signal distribution [2] [5] [9]*
Kurtosis	*The extent to which the distribution of signal amplitudes lies*

	predominantly on the left of the mean amplitude [2] [5] [9]
Average Power	*The variance per unit time [2] [9]*
Ratio of Spectral Peak (with Welch)	*Ratio of the energies of low and high frequency bands [2] [9]*
Signal Noise Ratio	*Power of whole signal over power of its computed noise [2]*
Total Harmonic Distortion	*Distortion of the whole signal compared to its harmonics [2]*

3. Methodology

This chapter describes how the dataset used in this thesis was gathered, what the dataset contains and details of how the dataset was processed. The steps described include our procedure for data collection, noise reduction, feature extraction and normalization. This processing flow is illustrated in figure 12 and is similar to the processing flow of gait verification/identification techniques presented in in [2] and [3]. Additionally, we added a normalization operation after the feature extraction step.

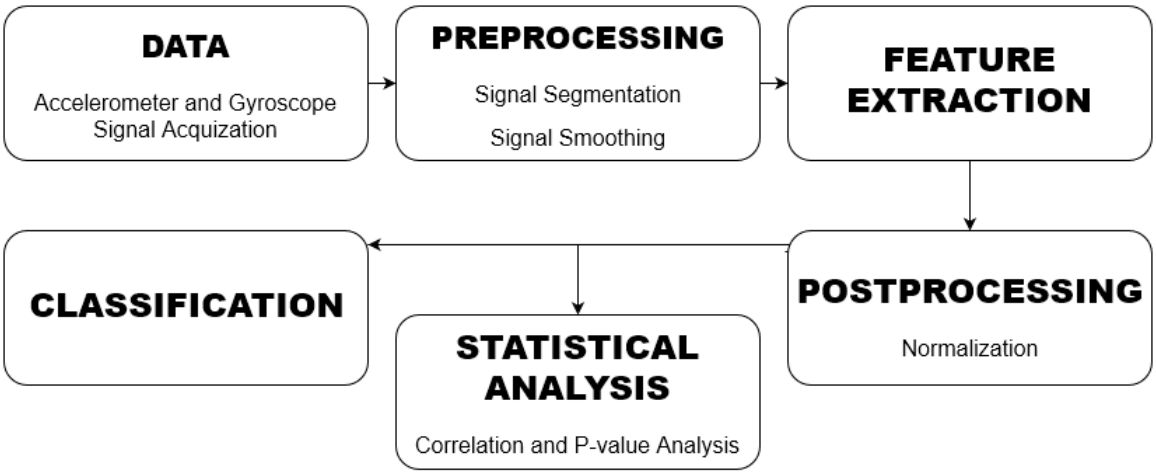

Figure 12 Work Flow of Signal Process and Analysis

Gait data was collected by Christina Aiello, a Masters Student in the WPI Computer Science Department, who is working on another aspect of passive detection of Alcohol Consumption from gait data. Details of the gait data will be described in Section 3.1. Pre-processing steps, segmentation and smoothing, are explained in Section 3.2. All feature extraction methods and functions used are described in Section 3.3. Our normalization method for accounting for differences in the walking patterns of different people will be described in Section 3.4. And Classification methods are described in Section3.5.

3.1 Data Collection and Dataset Summary

Christina Aiello gathered gait data from 24 people over a 5-week data-collection study using the MATLAB mobile data-collecting app. The data gathering app sampled the accelerometer and gyroscope in subjects' smartphones while they were walking. Intoxication is simulated in subjects by making them wear special goggles designed to simulate different BAC levels in the walks of subjects. The collected

data along with timestamps are saved into CSV files for further research. This thesis extracts signal processing features from the smartphone's accelerometer but does not process the gyroscope data in the dataset. Accelerometer data was represented by a set of x, y, z axis values.

Figure 13 Drunk Buster Goggles (left) and A User walking while wearing Drunk Buster Goggles [44]

Data from 9 of the subjects was excluded as their data was reliable, generated scattered readings or insufficient data to process.

All subjects walked while wearing drunk buster goggles rated at 0, 0.05, 0.12, 0.2 and 0.3 BAC alcohol levels. For each subject, 60 groups of sensor data was gathered from them. Each group consists of five segments, one for each of the BAC levels (0, 0.05, 0.12, 0.2 and 0.3) at which participant intoxication was simulated. Each segment lasted at least 5 seconds. Since the data was collected in segments, there was no need to segment it as part of the pre-process.

The following is a sample of accelerometer data. This sample shows a segment of 3.857 seconds accelerometer data from one person, when BAC value is 0. And entire data of a sample person can be found in the Appendix A: Data Samples.

Table 8 Data Sample of one person one segment of 3.857 seconds. Sampling under Approximately 10Hz. Related BAC = 0.

Accelerometer x (m/s^2)	Accelerometer y (m/s^2)	Accelerometer z (m/s^2)	Time stamp (s)
0.68354	-8.6592	-1.9279	0
1.4389	-5.7048	0.89364	0.09
120	-9.7743	-0.32082	0.189
48	-9.3374	-3.225	0.288
0	-9.3643	1.2821	0.388
5.7419	-13.969	2.7545	0.487
-4.5514	-2.8383	-1.5185	0.586

2.2116	-7.2197	-1.0313	0.685
-0.6608	-8.8149	-2.1805	0.784
0.18316	-13.743	-4.8836	0.883
2.2248	-5.7718	0.9517	0.982
1.5473	-9.7073	-1.7711	1.081
1.5227	-5.7569	1.3162	1.18
-0.56324	-10.655	-0.70449	1.279
0.098162	-11.241	-1.479	1.378
1.549	-11.808	5.3534	1.477
-0.58359	-16.654	1.4982	1.576
0.50578	-5.4779	-0.06045	1.675
0.73203	-7.6596	-2.0734	1.774
-1.5101	-10.862	-2.2817	1.873
1.2857	-17.183	0.68055	1.972
1.0223	-12.392	1.4856	2.071
1.4892	-6.0603	-2.7599	2.17
-1.6215	-10.491	-0.72664	2.269
0.07841	-8.8753	-0.85593	2.369
-1.3473	-12.253	-4.2066	2.482
-1.3216	-12.566	1.2983	2.582
-1.3216	-12.566	1.2983	2.667
1.2061	-4.9021	-1.0145	2.766
-0.2466	-11.314	-1.8417	2.865
-1.1756	-10.37	-3.7667	2.965
2.8784	-7.8835	3.2316	3.063
2.7988	-10.991	-0.02634	3.163
1.7615	-5.3852	1.1899	3.262
-0.65841	-10.633	-1.0463	3.362
-0.0826	-10.705	-1.6059	3.461
-0.90261	-9.0519	2.7186	3.56
1.8824	-18.872	0.083797	3.659
-1.2522	-4.1677	-0.26635	3.758
2.1817	-7.8703	-2.2075	3.857

3.2 Pre-processing

The pre-processing steps consist of segmentation at the beginning and a smoothing method to remove noise. Since the data was collected in 5-second segments, there was no need to segment the data. To smooth the data, a moving-average method was used to average out windows of accelerometer signals to reduce noise. The moving average calculation replaces each value in the sequence with the average of several points around it. We chose to average windows of 5 values, which balances both accuracy and time cost.

Since, SNR is one of our features, which relies on the noise, SNR was calculated before the

moving average was applied.

The following figure shows an example of the effect of moving average on a time sequence. The signal is smoothed after moving average is applied.

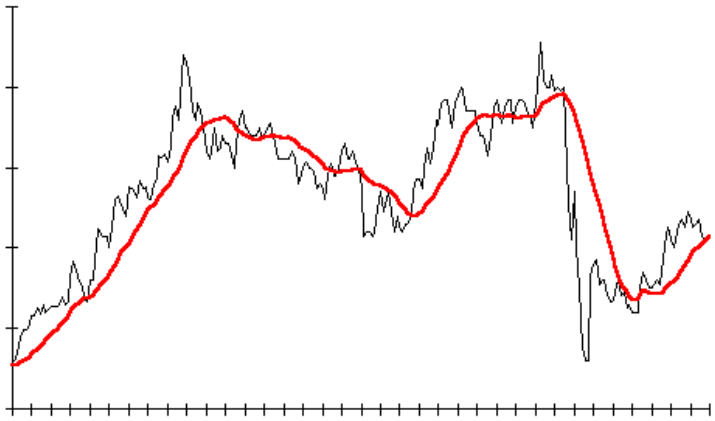

Figure 14 Example of Moving Average (in red) [64]

3.3 Feature extraction

In this thesis, the signal processing features investigated are generated in MATLAB functions (Code samples are in Appendix B):

Table 9 Table of MATLAB Function and Their Output Variables

Name of the Feature	MATLAB Function	Output Variables
Number of Steps	[ns, loc] = numSteps(x, y, z)	loc represents detected step locations. And ns represents number of steps in the sequence
Average Step Time	ast = averageStepTime(t, loc)	ast represents the average step time of this sequence
Average Cadence	ac = averageCadence(t, loc)	ac represents average cadence of this sequence,
Skewness	skew = skewness_acc(x, y, z)	skew represents the value of skewness of this sequence
Kurtosis	kurt = kurtosis_acc(x, y, z)	kurt represents the value of kurtosis of this sequence
Coefficient of Variation of Step Time	cvST = coef_var_stepTime(t, loc)	cvST represents coefficient of variation of step time of this sequence
Harmonic Ratio	hr = harmonicR(x, y, z)	hr represents the value of harmonic ratio of this sequence

Average Step Length	asl = averageStepLength(t, loc)	asl represents the average step length of this sequence
Gait Velocity	gv = gaitVelocity(t, loc)	gv represents the gait velocity of this sequence
Minimum and Maximum Difference	mmdiff = minMaxDiff(x, y, z)	mmdiff represents the values of minimum and maximum difference of this sequence
Standard Deviation	std_acc = std_acc(x, y, z)	std_acc represents the values of standard deviation of this sequence
Root Mean Square	rms_acc = rms_acc(x, y, z)	rms_acc represents the values of root mean square of this sequence
Entropy Rate	er = entropy_rate(x, y, z)	er represents the values of entropy rate of this sequence
Regression Line for Local Maxima and Minima	[minReg, maxReg] = regressionLineMaxMin(x, y, z)	minReg and maxReg represent a set of parameters describing the regression line of local minima and maxima of this sequence, respectively
Average Power	avg_pwr = averagePower(x, y, z)	avg_pwr represents the average power of this sequence
Ratio of Spectral Peak (with Welch)	rsp = ratioSpectralPeak(x, y, z)	rsp represents the ratio of spectral peak by welch
Ratio of Spectral Peak (with FFT)	rsp = ratioSpectralPeak_FFT(x, y, z)	rsp represents the ratio of spectral peak by fft
Ratio of Spectral Peak (with DCT)	rsp = ratioSpectralPeak_DCT(x, y, z)	rsp represents the ratio of spectral peak by dct
Signal Noise Ratio	snr_acc = snr_acc(x, y, z)	snr_acc represents signal noise ratio
Total Harmonic Distortion	thd_acc = thd_acc(x, y, z)	thd_acc represents the total harmonic distortion of this sequence
Energy in Band 0.5 to 3Hz	pFreq_05_3 = powerFreq_05_3(x, y, z)	pFreq_05_3 represents energy of frequency band from 0.5Hz to 3Hz
Windowed Energy in Band 0.5 to 3Hz	pFreq_05_3_w = powerFreq_05_3_windowed(x, y, z)	pFreq_05_3_w represents windowed energy of frequency band 0.5Hz to 3Hz
Peak Frequency	pFreq = peakFreq(x, y, z)	pFreq represents peak frequency of this sequence
Spectral Centroid	specC = spectralCentroid(x, y, z)	specC represents spectral centroid of this sequence
Bandwidth	bw = acc_bw(x, y, z)	bw represents bandwidth
Regression Line	pfReg =	pfReg represents a set of parameters

for windowed Energy	regressionLinePowerFreq_windowed(x, y, z)	describing the regression line of Windowed Energy in Band 0.5 to 3 Hz
Wavelet Bandwidth	wBW = wavelet_band(x, y, z)	wBW represents wavelet bandwidth
Wavelet Entropy Rate	wentropy = wavelet_entropy(x, y, z)	wentropy represents wavelet entropy rate
Zeroth-Lag Cross-Correlation Coefficient	r = cross_corr(x, y, z)	r represents cross correlation
Lampel-Ziv Complexit	lzc = lzComplexity(x, y, z)	lzc representsLampel-Ziv complexity

Table 10 Common Input Variable of MATLAB Functions

Input Variables	**Meaning**
x, y, z	accelerometer output in x-axis, y-axis, and z-axis, respectively
t	time
loc	detected step locations
ns	detected number of steps in the sequence

3.4 Normalization

After calculating features using their definition equations, normalization was applied to the features to account for variations in walking styles of various people. For example, people with different height may have different normal step length (figure below). And normalization will reduce such influence from the feature to get an accurate result.

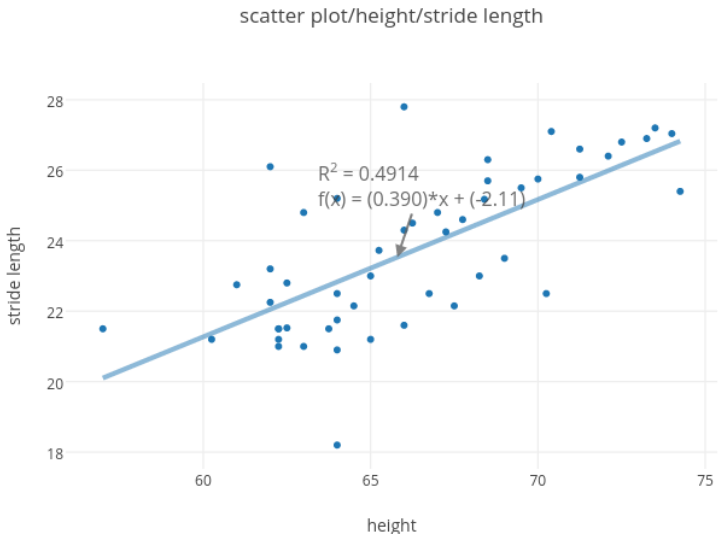

Figure 15 Individual's step length has influence on their normal step length

The normalization was done using the equation below.

$$Normalized\ vale = \frac{Raw\ Value}{Avarege\ Value\ of\ the\ same\ feature\ for\ the\ same\ person}$$

3.5 Classification

3.5.1 Classifiers

In the terminology of machine learning [35], classification is considered an instance of supervised learning. And an algorithm that implements classification, especially in a concrete implementation, is known as a classifier. There are multiple Classifiers in the field of machine learning. And the accuracy of them may vary according to different conditions. Thus, researchers tend to try and test different classifiers when they meet with specific classification problems.

In this thesis, I apply classification on the features with a p-value < 0.05 to prove that these features are not only promising, but also improve classification accuracy which is practically useful. I compared 5 popular classifiers: Random Forest, J48, JRip, NaiveBayes, and Decision Table. And the former two were also recommended by Arnold et al [2]. These classifier types are now briefly introduced.

Random Forest: Random forest is a notion of the general technique of random decision forests [36] that are an ensemble learning method for classification, regression and other tasks, which operate by constructing a multitude of decision trees at training time and outputting the class that is the mode

of the classes or mean prediction of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set [37]:587–588. The following figure illustrates the procedure of Random Forest.

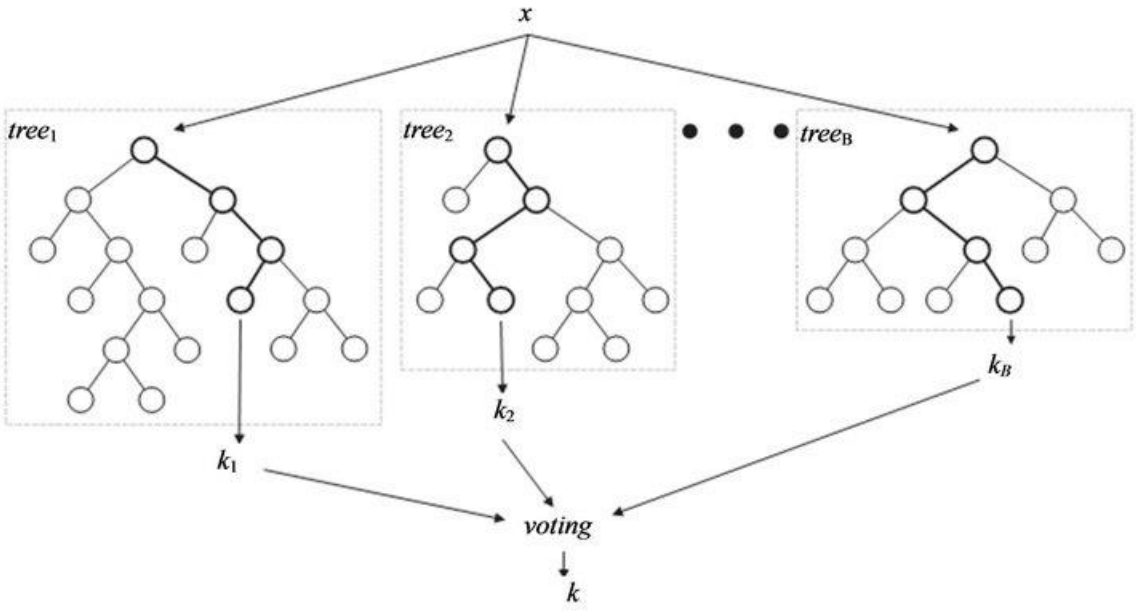

Figure 16 General Architecture of random forest [45]

J48: J48 is generating a pruned or unpruned C4.5 decision tree. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [38]. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often viewed as a statistical classifier [39].

When C4.5 algorithm is applied to a classification algorithm, a decision tree from the training data is generated using the concept of information entropy (Equation 1). Information entropy stands for the amount of information, which can help the decision tree to select those features that helps maximize the information entropy increase at their steps. [41]

$$H(V) = \sum_{k} P(v_k) \log_2 \frac{1}{P(v_k)}$$

Equation 1 Information Entropy, where v_k is a random variable

JRip: JRip refers to a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP [40]. In REP for rules algorithms, the training data is split into a growing set and a pruning set. Then the classifier

works in following steps:

An initial rule set is formed over the growing set, using some heuristic method, at the very beginning. The overlarge rule set is then simplified by applying one of a set of pruning operators repeatedly. Within each round of simplification, the pruning operator chosen is the one that yields the greatest reduction of error on the pruning set. When the moment comes that any pruning operator would increase error on the pruning set, not decreasing it as before, simplification can be viewed as complete [40].

Naive Bayes: Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem (Equation 2) with strong (naive) independence assumptions between the features [1].

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

Equation 2 Bayes' Theorem [46], $P(\theta|x)$ is the posterior probability of θ given predictor data x, and $P(x|\theta)$ is the likelihood of the predictor data given the class assignment

In WEKA, Naive Bayes Classifier is implemented using "Maximum A Posteriori" (MAP) rule. Under such rule, the prior probability distribution should be maximized [41], so the posterior probability density equation integrates to 1 (Equation 3).

$$\int_{\theta}^{argmax} P(x|\theta)P(\theta) = \int_{\theta}^{argmax} P(x)$$

Equation 3 MAP Rule for Naive Bayesian Networks [46]

Due to the fact that maximum-likelihood training can be done by evaluating a closed-form expression [41]:718, which takes linear time, Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. This property makes it stand out from the pool of classifiers.

SVM: Derived from Vapnik's statistical learning theory [47], Support Vector Machine (SVM) is a classifier of machine learning, for solving binary classification problems [48]. A binary classification problem refers to a classification task with only 2 classes (yes/no). A successful method to find Optimal Separating Hyperplane (OSH), which divides the group of data, is the key point to solve such problems.

As shown in the Figure below, SVM finds the OSH by maximizing the margin between 2 classes, on a higher dimensional transformed space. The points on the edge of the OSH margin are called Support Vectors, which support and hold the edge firmly [48]. In WEKA, the SVM method is implemented by using Sequential Minimal Optimization (SMO) algorithm from John Platt [2].

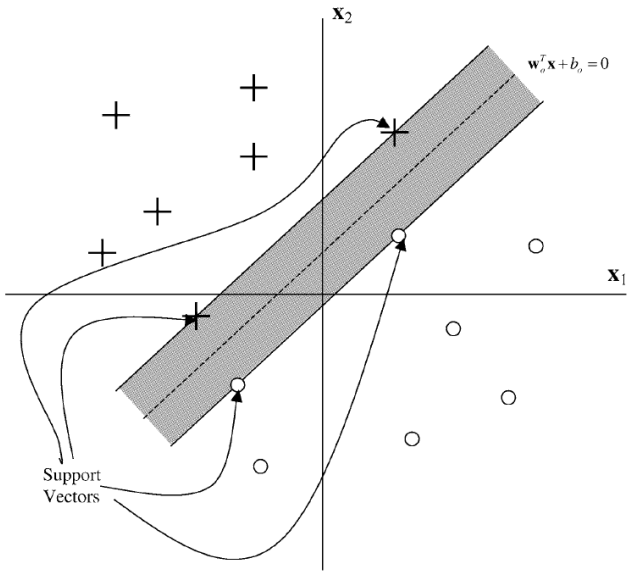

Figure 17 A Binary Classification Problem, with OSH (dash line marking $w_0{}^T x + b_0 = 0$) and Support Vectors. By mapping it to quadratic optimization problem with global minimum and linear constraints, an optimal w_0 and b_0 can be figured out [48]. Details of this calculation can be found in [49] [50]

Decision Table: Decision tables, like flowcharts and if-then-else and switch-case statements, associate conditions with actions to perform, but in many cases do so in a more elegant way. In a decision table, each decision corresponds to a variable, relation or predicate whose possible values are listed among the condition alternatives. Each action is a procedure or operation to perform, and the entries specify whether (or in what order) the action is to be performed for the set of condition alternatives the entry corresponds to. The following figure shows a simple example of decision table.

Printer troubleshooter

		Rules							
Conditions	Printer does not print	Y	Y	Y	Y	N	N	N	N
	A red light is flashing	Y	Y	N	N	Y	Y	N	N
	Printer is unrecognized	Y	N	Y	N	Y	N	Y	N
Actions	Check the power cable			X					
	Check the printer-computer cable	X		X					
	Ensure printer software is installed	X		X		X		X	
	Check/replace ink	X	X			X	X		
	Check for paper jam		X		X				

Figure 18 an Example of a Balanced Decision Table

3.5.2 Machine Learning Classifier Performance Metrics

In addition to classification accuracy, several performance metrics have been proposed for evaluating the results of machine learning algorithms. We now review the machine learning performance metrics used in our work.

Confusion Matrix: a confusion matrix shows the breakdown distribution of samples that are predicted correctly and incorrectly. Specifically, it gives a sense of what classes are being misclassified as what other classes. For a simple example, if we have 90 samples and 2 classes. The following table shows the structure a confusion matrix after a classifier is applied.

Table 11 Structure of a Confusion Matrix

	Classified as Class 1	Classified as Class 2
Class 1	Number of samples that is in Class 1 and also classified as Class 1	Number of samples that is in fact Class 1 but classified as Class 2
Class 2	Number of samples that is in fact Class 2 but classified as Class 1	Number of samples that is in Class 2 and also classified as Class 2

From confusion matrix, we can see the distribution of samples in terms of classes, and simplified it into true or false for a specific class. The following table shows the meaning of true positive, false positive, true negative and false negative. [2]

Table 12 True Positive, True Negative, False Positive and False Negative

	Nature True	Nature False
Prediction Positive	True Positive: predicted as true, and in fact true	False Positive: predicted as true, but in fact false
Prediction Negative	True Negative: predicted as false, but in fact true	False Negative: predicted as false, and in fact false

Thus, the equations for the *True Positive Rate and False Positive Rate:*

$$TP\ rate = \frac{TP}{TP + FN}$$

$$FP\ rate = \frac{FP}{FP + TN}$$

Where TP, FP, TN, FN refer to True Positives, False Positives, True Negatives and False Negatives respectively.

Precision:

$$Precision = \frac{TP}{TP + FP}$$

Recall:

$$Recall = \frac{TP}{TP + FN}$$

F-Measure:

$$F = 2 * \frac{precision * recall}{precision + recall}$$

ROC Area: ROC Area is the area under a Receiver Operating Characteristic (ROC) curve. The curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. When using normalized units, the area under the curve is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative') [42]. The following figure is an example of ROC curves.

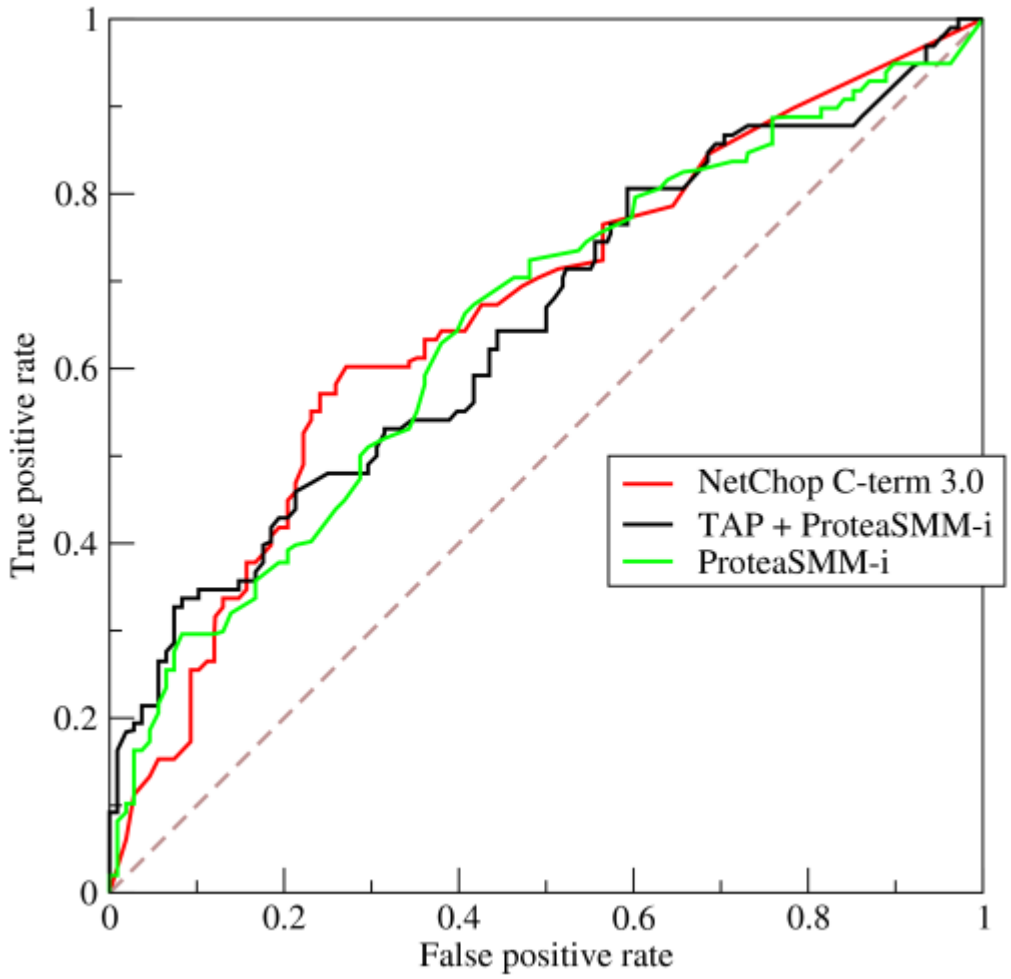

Figure 19 Example of ROC Curve

4. Results and Discussion

4.1 Effects of Normalizing Features

As is shown below in figures 14-20, most raw feature data is badly distributed with overlapping box plots (lower statistical significance), making correlation analysis less accurate. After normalization was applied, the distribution of the feature data generally became more compact with less overlap between adjacent boxplots, increasing statistical significance. Sample figures showing boxplots of features before and after normalization are shown in figures 14 - 20 below. More normalization comparison figures can be found in Appendix C: Normalization Results.

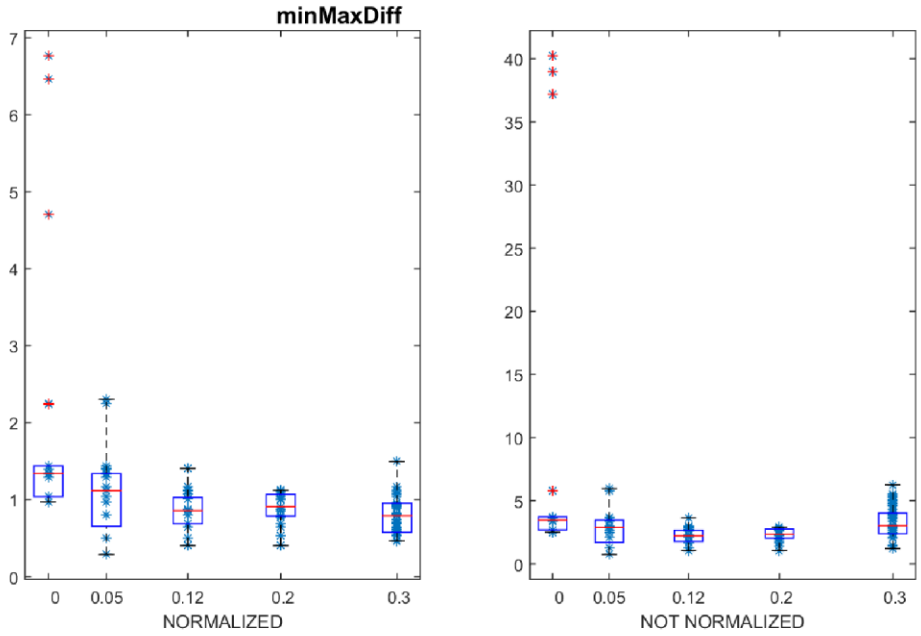

Figure 20 Data Distribution of Feature "Minimum and Maximum Difference" (Normalized on left vs. Not Normalized on right)

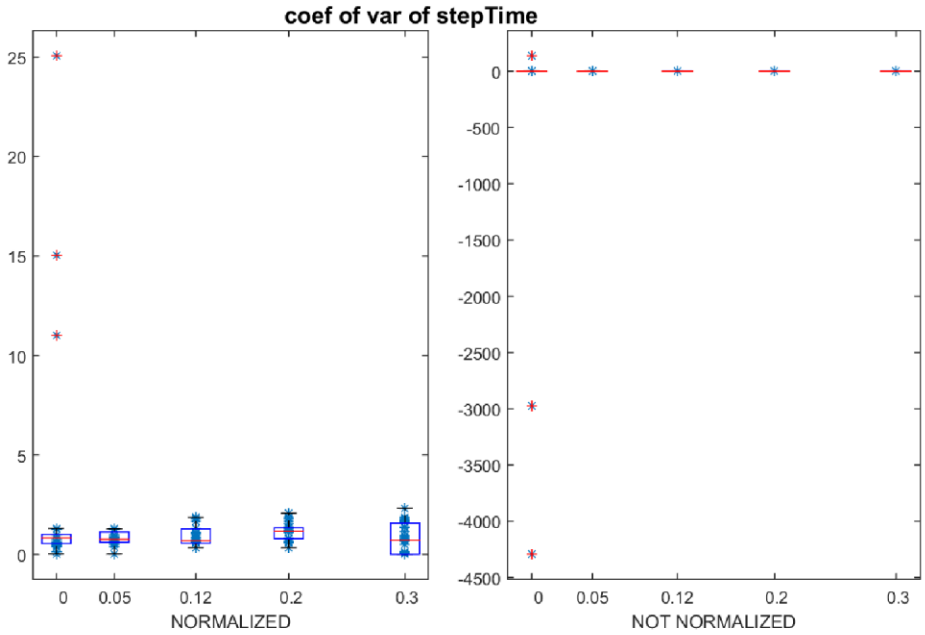

Figure 21 Data Distribution of Feature "Coefficient of Variation of Step Time" (Normalized on left vs. Not Normalized on right)

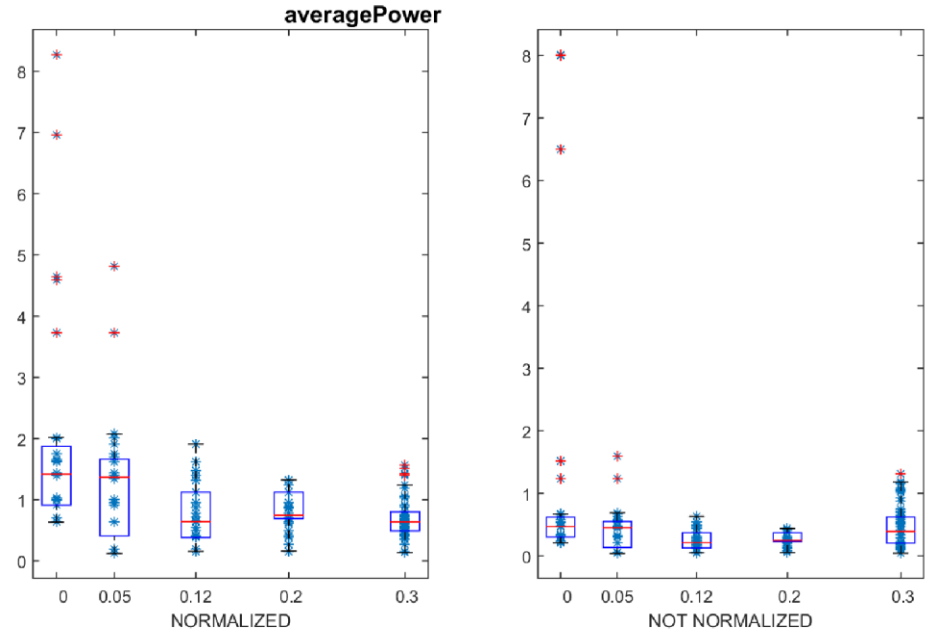

Figure 22 Data Distribution of Feature "Average Power" (Normalized on left vs. Not Normalized on right)

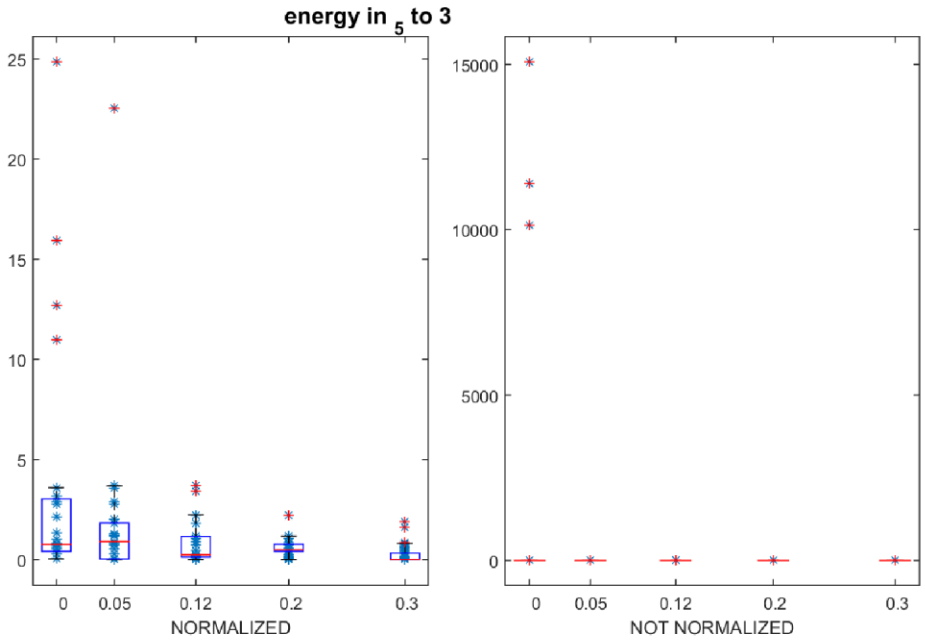

Figure 23 Data Distribution of Feature "Energy in Band 0.5 to 3 Hz" (Normalized on left vs. Not Normalized on right)

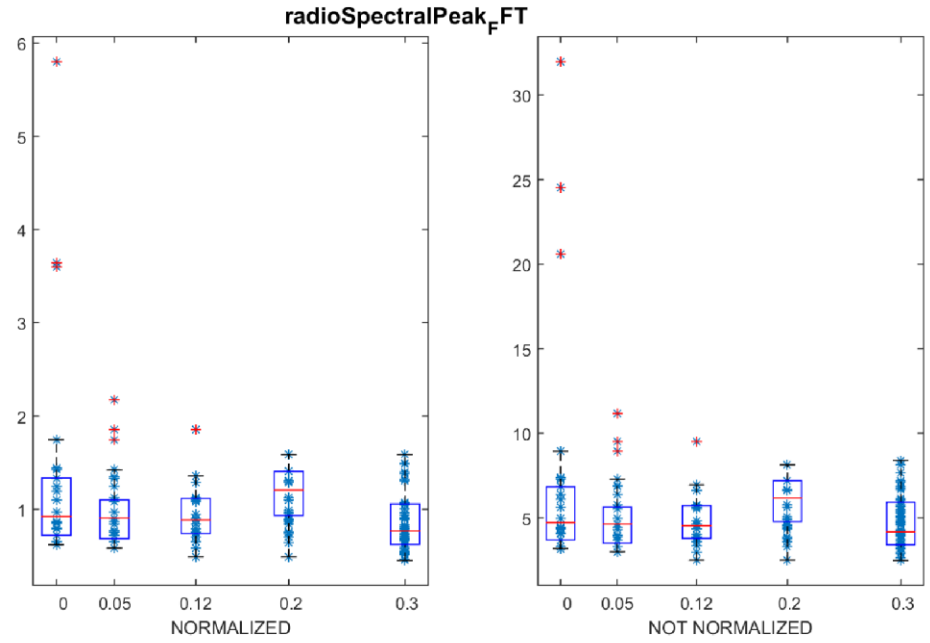

Figure 24 Data Distribution of Feature "Ratio of Spectral Peak by FFT" (Normalized on left vs. Not Normalized on right)

After showing the figure comparing the effect of normalization on each feature. I would also like to put two overall figures illustrating the distribution of feature value vs. BAC by boxplot.

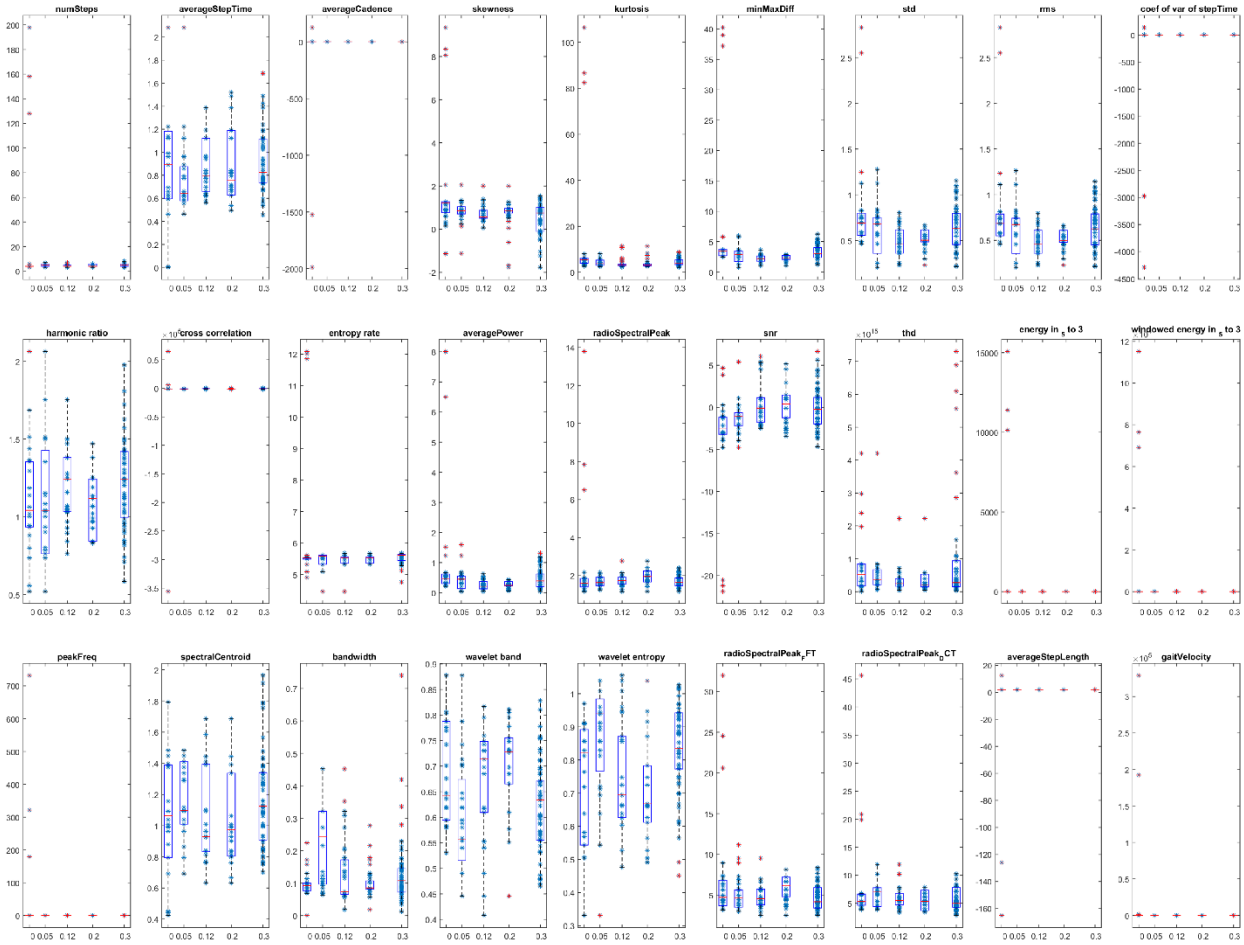

Figure 25 Boxplots of Features before Normalization

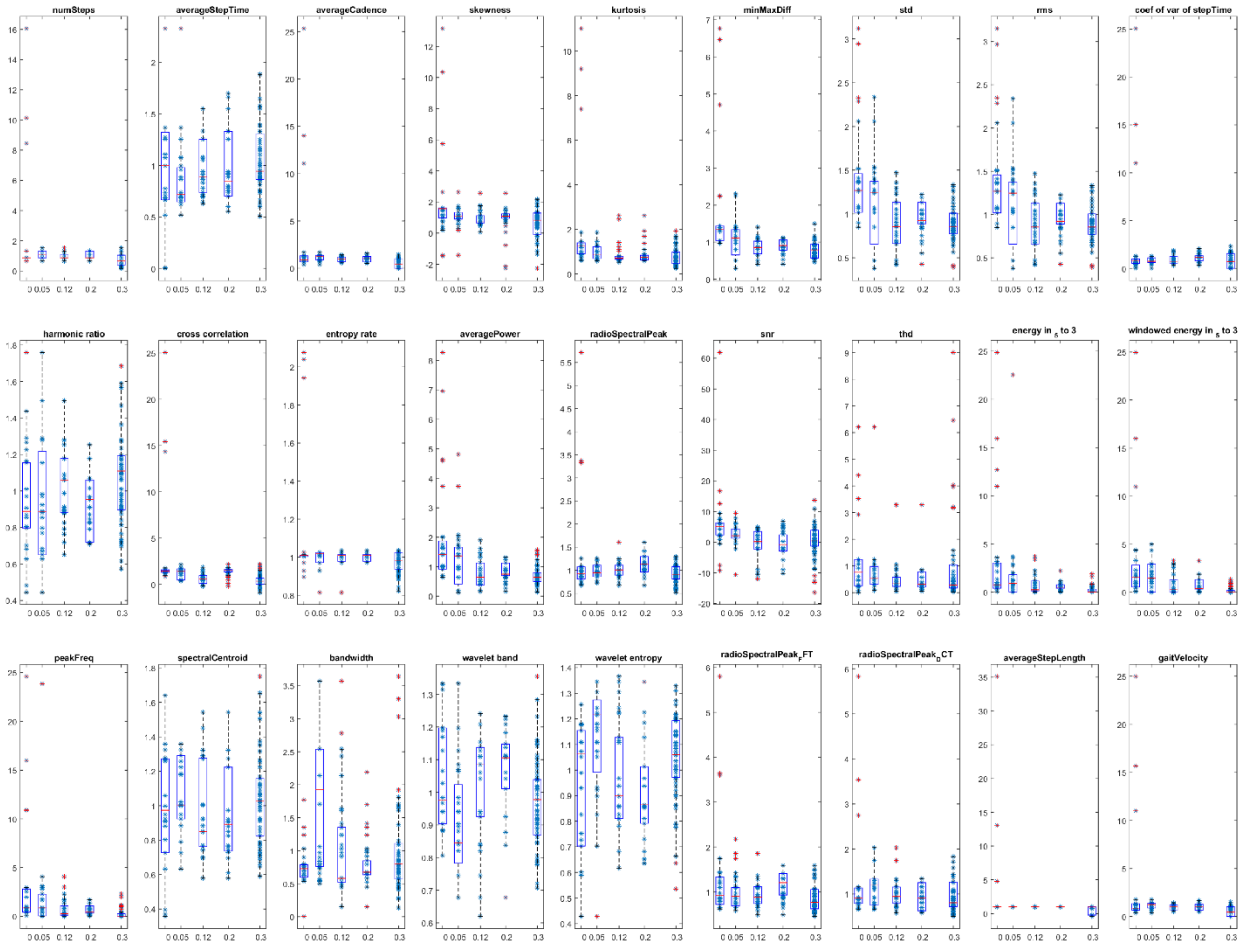

Figure 26 Boxplots of Features after Normalization

4.2 Correlation Based Feature Selection

In order to select the best features for gait classification, we utilized a Correlation based Feature Selection methodology (CFS) [34]. The Correlation Feature Selection (CFS) measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification (e.g. BAC levels), yet uncorrelated to each other.

For each class of features (e.g. time domain), we first calculate each feature's correlation with the labeled gait BAC levels, as well as their p-values. Features with p-values < 0.05 are useful for machine learning regardless of their correlation values. Hence, we filter out features with p-values greater than 0.05 and use all features with p-values < 0.05 as features in our supervised learning framework. This methodology is applied to all classes of features.

$$\rho(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\sigma_A}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right),$$

Equation 4 Correlation Coefficient

The figure below [63] shows the definition of P-value. So if the p-value is lower than 0.05, the more likely observation covers over 95%, indicating a positive response as a feature in machine learning.

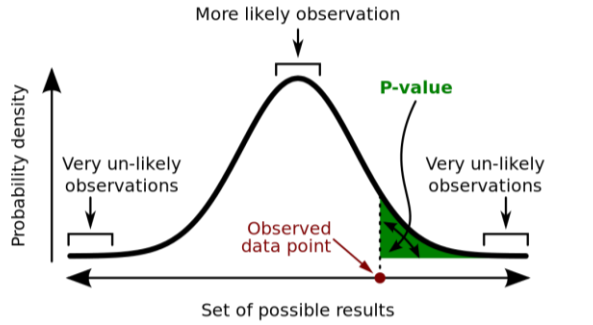

Figure 27 Definition of P-value

4.2.1 Time Domain Features and Ranking

Table 12 shows time domain features with p-value < 0.05 and ranked based on their correlation values. As shown in the table, 12 out of 13 time domain features had p-value < 0.05 and were potentially useful in classifying alcohol consumption detection. Normalization further increased the

correlation of all 12 features by an average of 0.1061.

Table 13 Time Domain Features Ranked by Correlation Coefficient

		Before Normalization			After Normalization			
Index	Feature Names	Features Coef	P-value	Predictable (p<0.05)	Features Coef	P-value	Predictable (p<0.05)	Coef Diff
1	**std**	**-0.1068**	**0.0657**	**0**	**-0.3947**	**0.0000**	**1**	**0.2880**
2	**rms**	**-0.1067**	**0.0660**	**0**	**-0.3943**	**0.0000**	**1**	**0.2877**
3	**minMaxDiff**	**-0.1268**	**0.0286**	**1**	**-0.3842**	**0.0000**	**1**	**0.2574**
4	**skewness**	**-0.2649**	**0.0000**	**1**	**-0.2715**	**0.0000**	**1**	**0.0066**
5	**kurtosis**	**-0.1509**	**0.0091**	**1**	**-0.2610**	**0.0000**	**1**	**0.1101**
6	**gaitVelocity**	**-0.1131**	**0.0511**	**0**	**-0.2523**	**0.0000**	**1**	**0.1392**
7	**averageCadence**	**0.1108**	**0.0561**	**0**	**-0.2490**	**0.0000**	**1**	**0.1383**
8	**numSteps**	**-0.1309**	**0.0238**	**1**	**-0.2102**	**0.0003**	**1**	**0.0793**
9	**averageStepLength**	**0.1108**	**0.0561**	**0**	**-0.1988**	**0.0006**	**1**	**0.0880**
10	**entropy rate**	**-0.0773**	**0.1831**	**0**	**-0.1813**	**0.0017**	**1**	**0.1040**
11	**harmonic ratio**	**0.1505**	**0.0093**	**1**	**0.1708**	**0.0031**	**1**	**0.0203**
12	**coef of var of stepTime**	**0.1128**	**0.0518**	**0**	**-0.1346**	**0.0202**	**1**	**0.0218**
	Average Useful	0.1302			0.2586			0.1284
13	*averageStepTime*	*0.0831*	*0.1525*	*0*	*0.0975*	*0.0928*	*0*	*0.0000*
	Average All	0.1251			0.2312			0.1061

Then the 12 features with p-value < 0.05 were classified using the WEKA machine learning library using 10-fold cross-validation. The accuracy of different classifiers are listed below. The most accurate classifier type is Random Forest with an accuracy of 83.22%.

Table 14 Classifiers Ranked by Accuracy for Time Domain features with p-value < 0.05

Classifier Type	Accuracy
RandomForest	83.22%
JRip	80.20%
J48	78.86%
DecisionTable	74.16%
NaiveBayes	48.66%
SMO (SVM in WEKA)	41.28%

The confusion matrix of the Random Forest classifier is shown in table 7 below. TP Rate, FP Rate, precision, recall, F-measure and ROC area are reported in table 14. The confusion matrix describes the correct and confused classifications in detail. For example the first row of data in confusion matrix shows that 49 samples of BAC = 0 were classified as BAC = 0, which are correct. And 2 samples of BAC =

0 are mis-classified as BAC = 0.05, which is wrong.

Table 15 Classification Performance Metrics for Time Domain features with p-value < 0.05

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BAC==0	0.942	0.049	0.803	0.942	0.867	0.968
BAC==0.05	0.625	0.039	0.714	0.625	0.667	0.855
BAC==0.12	0.807	0.054	0.780	0.807	0.793	0.909
BAC==0.2	0.632	0.035	0.727	0.632	0.676	0.836
BAC==0.3	0.937	0.032	0.945	0.937	0.941	0.979
Weighted Avg.	0.932	0.040	0.830	0.832	0.829	0.929

Table 16 Confusion Matrix for Time Domain features with p-value < 0.05

BAC=0	BAC=0.05	BAC=0.12	BAC=0.2	BAC=0.3	<-classified as
49	2	0	1	0	BAC==0
10	25	4	1	0	BAC==0.05
1	6	46	3	1	BAC==0.12
0	0	9	24	5	BAC==0.2
1	2	0	4	104	BAC==0.3

4.2.2 Frequency Domain Features and Ranking

Table 16 shows frequency domain features with p-value < 0.05 and ranked based on their correlation values. As shown in the table, 8 out of 11 features were statistically significant (p-value < 0.05) and were potentially useful in alcohol consumption detection. 7 out of these 8 frequency domain features showed stronger correlation after normalization by an average of 0.0999.

The Regression Line of Windowed Energy feature was excluded since we were unable to achieve a reasonable implementation.

Table 17 Frequency Domain Features Ranked by Correlation Coefficient

		Before Normalization			After Normalization			
Index	Feature Names	Features Coef	P-value	Predictable (p<0.05)	Features Coef	P-value	Predictable (p<0.05)	Coef Diff
1	**averagePower**	**-0.1345**	**0.0202**	**1**	**-0.3990**	**0.0000**	**1**	**0.2645**
2	**windowed energy in _5 to 3**	**-0.1393**	**0.0161**	**1**	**-0.3974**	**0.0000**	**1**	**0.2581**
3	**energy in _5 to 3**	**-0.1409**	**0.0149**	**1**	**-0.3347**	**0.0000**	**1**	**0.1937**
4	**peakFreq**	**-0.1239**	**0.0325**	**1**	**-0.3196**	**0.0000**	**1**	**0.1958**
5	**snr**	**0.2669**	**0.0000**	**1**	**-0.2471**	**0.0000**	**1**	**-0.0199**
6	**ratioSpectralPeak_FFT**	**-0.1385**	**0.0168**	**1**	**-0.1734**	**0.0027**	**1**	**0.0349**
7	**ratioSpectralPeak**	**-0.0925**	**0.1111**	**0**	**-0.1703**	**0.0032**	**1**	**0.0778**
8	**ratioSpectralPeak_DCT**	**-0.1179**	**0.0420**	**1**	**-0.1525**	**0.0084**	**1**	**0.0346**
	Average Useful	0.1443			0.2742			0.1299
9	*bandwidth*	*-0.0682*	*0.2408*	*0*	*-0.0795*	*0.1711*	*0*	*0.0000*
10	*spectralCentroid*	*0.0910*	*0.1168*	*0*	*0.0393*	*0.4996*	*0*	*0.0000*
11	*thd*	*0.1056*	*0.0687*	*0*	*0.0362*	*0.5334*	*0*	*0.0000*
	Average All	0.1314			0.2313			0.0999

Then the 8 features with p-value < 0.05 were classified using the WEKA machine learning library using 10-fold cross-validation. The accuracy of different classifiers are listed in table 17 below. The most accurate classifier type is J48 with an accuracy of 82.21%.

Table 18 Classifiers Ranked by Accuracy for Frequency Domain features with p-value < 0.05

Classifier Type	Accuracy
J48	82.21%
RandomForest	79.53%
JRip	77.18%
DecisionTable	74.83%
NaiveBayes	48.99%

SMO (SVM in WEKA)	43.29%

The confusion matrix of the J48 classifier is shown in table 18 below. TP Rate, FP Rate, precision, recall, F-measure and ROC area are reported in table 10. The confusion matrix describes the correct and confused classifications in detail. For example the first row of data in the confusion matrix shows that 46 samples of BAC = 0 are classified as BAC = 0, which are correct. And 2 samples of BAC = 0 are mis-classified as BAC = 0.05, which is wrong.

Table 19 Detailed Accuracy for Frequency Domain features with p-value < 0.05

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BAC==0	0.885	0.061	0.754	0.885	0.814	0.913
BAC==0.05	0.650	0.027	0.788	0.650	0.712	0.904
BAC==0.12	0.842	0.079	0.716	0.842	0.774	0.874
BAC==0.2	0.605	0.023	0.793	0.605	0.687	0.852
BAC==0.3	0.919	0.032	0.944	0.919	0.932	0.958
Weighted Avg.	0.822	0.044	0.827	0.822	0.820	0.913

Table 20 Confusion Matrix for Frequency Domain features with p-value < 0.05

BAC=0	BAC=0.05	BAC=0.12	BAC=0.2	BAC=0.3	<-classified as
46	2	3	0	1	BAC==0
13	26	1	0	0	BAC==0.05
0	5	48	1	3	BAC==0.12
1	0	12	23	2	BAC==0.2
1	0	3	5	102	BAC==0.3

4.2.3 Wavelet Domain Features and Ranking

Table 20 shows wavelet domain features with p-value < 0.05 and ranked based on their correlation values. As shown in the table below, 1 of the 2 features were useful in alcohol consumption detection.

Normalization does not improve the performance of wavelet domain features. This condition probably results from the property of the wavelet domain. Wavelet domain is a time-frequency domain, which reflects not only time and frequency properties, but also the relationship between time and frequency. However, the normalization process, which usually resizes the range of feature values, may reshape the relationship between time and frequency, causing a decrease in the feature correlation coefficient.

Table 21 Wavelet Domain Features and Ranking by Correlation Coefficient

		Before Normalization			After Normalization			
Index	Feature Names	Features Coef	P-value	Predictable (p<0.05)	Features Coef	P-value	Predictable (p<0.05)	Coef Diff
1	**wavelet entropy**	**0.1880**	**0.0011**	**1**	**0.1229**	**0.0340**	**1**	**-0.0651**
	Average Useful	0.1880			0.1229			-0.0651
2	*wavelet band*	*-0.1565*	*0.0068*	*1*	*-0.0889*	*0.1256*	*0*	*0.0000*
	Average All	0.1723			0.1059			-0.0664

Then the feature with p-value < 0.05 was classified using the WEKA machine learning library using 10-fold cross-validation. The accuracy of different classifiers are listed below. The most accurate classifier type is Random Forest with an accuracy of 77.85%.

Table 22 Classifiers Ranked by Accuracy for Wavelet Domain features with p-value < 0.05

Classifier Type	**Accuracy**
RandomForest	77.85%
J48	75.84%
JRip	70.81%
DecisionTable	53.36%
NaiveBayes	42.62%
SMO (SVM in WEKA)	37.25%

The confusion matrix of the Random Forest classifier is shown in table 22 below. TP Rate, FP Rate, precision, recall, F-measure and ROC area are reported in table 14. The confusion matrix describes the correct and confused classifications in detail. For example the first row of data in confusion matrix

shows that 45 samples of BAC = 0 are classified as BAC = 0, which are correct. And 3 samples of BAC = 0 are mis-classified as BAC = 0.05, which is wrong.

Table 23 Detailed Accuracy for Wavelet Domain features with p-value < 0.05

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BAC==0	0.865	0.073	0.714	0.865	0.783	0.905
BAC==0.05	0.600	0.054	0.632	0.600	0.615	0.740
BAC==0.12	0.807	0.066	0.742	0.807	0.773	0.879
BAC==0.2	0.658	0.038	0.714	0.658	0.685	0.789
BAC==0.3	0.829	0.043	0.920	0.829	0.872	0.910
Weighted Avg.	0.779	0.054	0.785	0.779	0.779	0.865

Table 24 Confusion Matrix for Wavelet Domain features with p-value < 0.05

BAC=0	BAC=0.05	BAC=0.12	BAC=0.2	BAC=0.3	<-classified as
45	3	0	1	3	BAC==0
10	24	4	0	2	BAC==0.05
3	5	46	2	1	BAC==0.12
2	0	9	25	2	BAC==0.2
3	6	3	7	92	BAC==0.3

4.2.4 Statistical Features and Ranking

Table 24 shows statistical domain features with p-value < 0.05 and ranked based on their correlation values. As shown in tables 16 below, all 3 features had a p-value < 0.05 and were potentially useful in alcohol consumption detection. Additionally, all 3 features showed stronger correlation after normalization.

Table 25 Statistical Features and Ranking by Correlation Coefficient

Index	Feature Names	Features Coef	P-value	Predictable (p<0.05)	Features Coef	P-value	Predictable (p<0.05)	Coef Diff
7	**std**	**-0.1068**	**0.0657**	**0**	**-0.3947**	**0.0000**	**1**	**0.2880**
11	**cross correlation**	**0.0720**	**0.2152**	**0**	**-0.2848**	**0.0000**	**1**	**0.2128**
5	**kurtosis**	**-0.1509**	**0.0091**	**1**	**-0.2610**	**0.0000**	**1**	**0.1101**
	Average	0.1099			0.3135			0.2036

Then the 3 features with p-value < 0.05 were classified in the WEKA machine learning library using 10-fold cross-validation. The accuracy of different classifiers are listed below. The most accurate classifier type is J48 with an accuracy of 83.89%.

Table 26 Classifiers Ranked by Accuracy for Statistical features with p-value < 0.05

Classifier Type	**Accuracy**
J48	83.89%
RandomForest	82.86%
JRip	76.51%
DecisionTable	72.15%
NaiveBayes	50.34%
SMO (SVM in WEKA)	40.94%

The confusion matrix of the J48 classifier is shown in table 26 below. TP Rate, FP Rate, precision, recall, F-measure and ROC area are reported in table 18. The confusion matrix describes the correct and confused classifications in detail. For example the first row of data in confusion matrix shows that 47 sample of BAC = 0 are classified as BAC = 0, which are correct. And 2 samples of BAC = 0 are mis-classified as BAC = 0.05, which is wrong.

Table 27 Detailed Accuracy for Statistical features with p-value < 0.05

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
BAC==0	0.904	0.041	0.825	0.904	0.862	0.942
BAC==0.05	0.775	0.027	0.816	0.775	0.795	0.908
BAC==0.12	0.772	0.058	0.759	0.772	0.765	0.874
BAC==0.2	0.632	0.038	0.706	0.632	0.667	0.826

BAC==0.3	0.937	0.037	0.937	0.937	0.937	0.977
Weighted Avg.	0.839	0.041	0.837	0.839	0.839	0.923

Table 28 Confusion Matrix for Statistical features with p-value < 0.05

BAC=0	**BAC=0.05**	**BAC=0.12**	**BAC=0.2**	**BAC=0.3**	**<-classified as**
47	2	1	0	2	BAC==0
7	31	0	0	2	BAC==0.05
1	5	44	4	3	BAC==0.12
2	0	12	24	0	BAC==0.2
0	0	1	6	104	BAC==0.3

4.2.5 Information-Theoretic Features and Ranking

Table 28 shows statistical domain features with p-value < 0.05 and ranked based on their correlation values. As shown in the table below, 1 feature had a p-value < 0.05 and was useful in alcohol consumption detection. And this feature showed stronger correlation with the BAC levels after normalization. We were unable to implement the the "Lempel-Ziv Complexity" feature in a reasonable time, so we excluded it.

Table 29 Information-Theoretic Features and Ranking by Correlation Coefficient

Index	Feature Names	Features Coef	P-value	Predictable (p<0.05)	Features Coef	P-value	Predictable (p<0.05)	Coef Diff
1	**entropy rate**	**-0.0773**	**0.1831**	**0**	**-0.1813**	**0.0017**	**1**	**0.1040**
	Average	0.0773			0.1813			0.1040

Then the Information-Theoretic feature with p-value < 0.05 was classified using the WEKA machine learning library using 10-fold cross-validation. The accuracy of different classifiers are listed below. The most accurate classifier type is Random Forest with an accuracy of 58.05%.

Table 30 Classifiers Ranked by Accuracy for Information-Theoretic features with p-value < 0.05

Classifier Type	Accuracy
RandomForest	58.05%
J48	57.05%
DecisionTable	53.36%
JRip	43.29%
NaiveBayes	37.92%
SMO (SVM in WEKA)	37.25%

The confusion matrix of the Random Forest classifier is shown in table 30 below. TP Rate, FP Rate, precision, recall, F-measure and ROC area are reported in table 22. The confusion matrix describes the correct and confused classifications in detail. For example the first row of data in confusion matrix shows that 34 samples of BAC = 0 are classified as BAC = 0, which are correct. And 4 samples of BAC = 0 are mis-classified as BAC = 0.05, which is wrong.

Table 31 Detailed Accuracy for Information-Theoretic features with p-value < 0.05

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BAC==0	0.654	0.195	0.415	0.654	0.507	0.820
BAC==0.05	0.650	0.097	0.510	0.650	0.571	0.776
BAC==0.12	0.298	0.100	0.415	0.298	0.347	0.768

BAC==0.2	0.237	0.031	0.529	0.237	0.327	0.709
BAC==0.3	0.784	0.107	0.813	0.784	0.798	0.901
Weighted Avg.	0.581	0.110	0.590	0.581	0.571	0.820

Table 32 Confusion Matrix for Information-Theoretic features with p-value < 0.05

BAC=0	BAC=0.05	BAC=0.12	BAC=0.2	BAC=0.3	<-classified as
34	4	9	0	5	BAC==0
9	26	3	0	2	BAC==0.05
20	4	17	7	9	BAC==0.12
13	6	6	9	4	BAC==0.2
6	11	6	1	87	BAC==0.3

4.2.6 All Useful Features and Ranking

The following Table lists and ranks all 22 features with p-value < 0.05 ranked by correlation coefficient. 20 out of the 22 useful features showed stronger correlation with BAC levels after Normalization was applied.

Table 33 All Useful Features and Ranking by Correlation Coefficient

		Before Normalization			After Normalization			
Index	Feature Names	Features Coef	P-value	Predictable (p<0.05)	Features Coef	P-value	Predictable (p<0.05)	Coef Diff
1	**averagePower**	**-0.1345**	**0.0202**	**1**	**-0.3990**	**0.0000**	**1**	**0.2645**
2	**windowed energy in _5 to 3**	**-0.1393**	**0.0161**	**1**	**-0.3974**	**0.0000**	**1**	**0.2581**
3	**Std**	**-0.1068**	**0.0657**	**0**	**-0.3947**	**0.0000**	**1**	**0.2880**
4	**Rms**	**-0.1067**	**0.0660**	**0**	**-0.3943**	**0.0000**	**1**	**0.2877**
5	**minMaxDiff**	**-0.1268**	**0.0286**	**1**	**-0.3842**	**0.0000**	**1**	**0.2574**
6	**energy in _5 to 3**	**-0.1409**	**0.0149**	**1**	**-0.3347**	**0.0000**	**1**	**0.1937**
7	**peakFreq**	**-0.1239**	**0.0325**	**1**	**-0.3196**	**0.0000**	**1**	**0.1958**
8	**cross correlation**	**0.0720**	**0.2152**	**0**	**-0.2848**	**0.0000**	**1**	**0.2128**
9	**Skewness**	**-0.2649**	**0.0000**	**1**	**-0.2715**	**0.0000**	**1**	**0.0066**
10	**Kurtosis**	**-0.1509**	**0.0091**	**1**	**-0.2610**	**0.0000**	**1**	**0.1101**
11	**gaitVelocity**	**-0.1131**	**0.0511**	**0**	**-0.2523**	**0.0000**	**1**	**0.1392**
12	**averageCadence**	**0.1108**	**0.0561**	**0**	**-0.2490**	**0.0000**	**1**	**0.1383**
13	**Snr**	**0.2669**	**0.0000**	**1**	**-0.2471**	**0.0000**	**1**	**-0.0199**
14	**numSteps**	**-0.1309**	**0.0238**	**1**	**-0.2102**	**0.0003**	**1**	**0.0793**
15	**averageStepLength**	**0.1108**	**0.0561**	**0**	**-0.1988**	**0.0006**	**1**	**0.0880**
16	**entropy rate**	**-0.0773**	**0.1831**	**0**	**-0.1813**	**0.0017**	**1**	**0.1040**
17	**ratioSpectralPeak_FFT**	**-0.1385**	**0.0168**	**1**	**-0.1734**	**0.0027**	**1**	**0.0349**
18	**harmonic ratio**	**0.1505**	**0.0093**	**1**	**0.1708**	**0.0031**	**1**	**0.0203**
19	**ratioSpectralPeak**	**-0.0925**	**0.1111**	**0**	**-0.1703**	**0.0032**	**1**	**0.0778**
20	**ratioSpectralPeak_DCT**	**-0.1179**	**0.0420**	**1**	**-0.1525**	**0.0084**	**1**	**0.0346**
21	**coef of var of stepTime**	**0.1128**	**0.0518**	**0**	**-0.1346**	**0.0202**	**1**	**0.0218**
22	**wavelet entropy**	**0.1880**	**0.0011**	**1**	**0.1229**	**0.0340**	**1**	**-0.0651**
	Number of Useful			14			22	
	Average Useful	0.1353			0.2593			0.1240

These ranked list of features with p-value < 0.05 are:

1. Average Power
2. Windowed Energy in Band 0.5 to 3 Hz
3. Standard Deviation

4. Root Mean Square
5. Minimum and Maximum Difference
6. Energy in Band 0.5 to 3 Hz
7. Peak Frequency
8. Zero-Lag Cross-Correlation Coefficient
9. Skewness
10. Kurtosis
11. Gait Velocity
12. Average Cadence
13. Signal Noise Ratio
14. Number of Steps
15. Average Step Length
16. Entropy Rate
17. Ratio of Spectral Peak by FFT
18. Harmonic Ratio
19. Ratio of Spectral Peak
20. Ratio of Spectral Peak by DCT
21. Coefficient of Variation of Step Time
22. Wavelet Entropy

These 22 features with p-value < 0.05 were used in supervised classification of the gait BAC levels in WEKA with 10-fold cross-validation, yielding an accuracy of 84.90% using Random Forest Classifier.

Table 34 Classifiers Ranked by Accuracy for features with p-value < 0.05

Classifier Type	Accuracy
RandomForest	84.90%
J48	80.87%
JRip	80.54%
DecisionTable	75.17%
NaiveBayes	56.04%
SMO (SVM in WEKA)	43.62%

The confusion matrix of the Random Forest classifier is shown in table 34 below. TP Rate, FP Rate, precision, recall, F-measure and ROC area are reported in table 26. The confusion matrix describes

the correct and confused classifications in detail. For example the first row of data in confusion matrix shows that 49 sample of BAC = 0 are classified as BAC = 0, which are correct. And 2 samples of BAC = 0 are classified as BAC = 0.05, which is confused.

Table 35 Detailed Accuracy for features with p-value < 0.05

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
BAC==0	0.942	0.041	0.031	0.942	0.883	0.969
BAC==0.05	0.650	0.031	0.765	0.650	0.703	0.7854
BAC==0.12	0.825	0.054	0.783	0.825	0.803	0.906
BAC==0.2	0.7121	0.042	0.711	0.7121	0.711	0.848
BAC==0.3	0.937	0.016	0.972	0.937	0.954	0.974
Weighted Avg.	0.849	0.033	0.850	0.849	0.848	0.928

Table 36 Confusion Matrix for features with p-value < 0.05

BAC=0	BAC=0.05	BAC=0.12	BAC=0.2	BAC=0.3	<-classified as
49	2	0	0	1	BAC==0
10	26	4	0	0	BAC==0.05
0	6	47	4	0	BAC==0.12
0	0	9	27	2	BAC==0.2
0	0	0	7	104	BAC==0.3

5. Conclusion

Based on the results presented in the previous chapter, several conclusions can be made in conclusion.

1. As we can see in the boxplot and predictability report, normalization improves the performance of most (20 out of 22) features that had p-values < 0.05.
2. Statistical Features has the best accuracy of 83.89%. Time Domain Feature and Frequency Domain Features follow with accuracies of 83.22% and 82.21%, respectively.
3. Frequency domain features may be improved by using Time-Frequency Transform methods, as preliminary experiments show that the p-value changes when we switch between Welch, FFT and DCT for calculating the ratio of Spectral Peaks.
4. There are 22 features among 27 tested features are promising for prediction. A ranked list of these features can be referred in section 4.2.6.

6. Future Work

Two features that we planned to implement could not be achieved in a reasonable time. Hence we now list them as future work. These features that will be implemented in future are:

1. Lempel-Ziv Complexity.
2. Regression Line for Local Maxima and Minima, which requires walked distance to be calculated.
3. Regression Lines for Windowed Energy, which requires walked distance to be calculated. .

Bibliography

1. Arnold, Z., Larose, D., & Agu, E. (2015, October). Smartphone Inference of Alcohol Consumption Levels from Gait. In Healthcare Informatics (ICHI), 2015 International Conference on (pp. 417-426). IEEE.
2. Arnold, Z. (2015). *Smartphone Gait Inference* (Doctoral dissertation, WORCESTER POLYTECHNIC INSTITUTE).
3. Derawi, M. O. (2010). Accelerometer-based gait analysis, a survey. *Nor Informasjonssikkerhetskonferanse NISK*.
4. Klucken, J., Barth, J., Kugler, P., Schlachetzki, J., Henze, T., Marxreiter, F., ... & Winkler, J. (2013). Unbiased and mobile gait analysis detects motor impairment in Parkinson's disease. PloS one, 8(2), e56956.
5. Sejdic, E., Lowry, K. A., Bellanca, J., Redfern, M. S., & Brach, J. S. (2014). A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, *22*(3), 603-612.
6. Derawi, M. O., Bours, P., & Holien, K. (2010, October). Improved cycle detection for accelerometer based gait authentication. In *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on* (pp. 312-317). IEEE.
7. Cho, D. K., Mun, M., Lee, U., Kaiser, W. J., & Gerla, M. (2010, March). Autogait: A mobile platform that accurately estimates the distance walked. In *Pervasive computing and communications (PerCom), 2010 IEEE international conference on* (pp. 116-124). IEEE.
8. Wittlinger, M., Wehner, R., & Wolf, H. (2007). The desert ant odometer: a stride integrator that accounts for stride length and walking speed. *Journal of experimental Biology*, *210*(2), 198-207.
9. Lu, H., Huang, J., Saha, T., & Nachman, L. (2014, September). Unobtrusive gait verification for mobile phones. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers* (pp. 91-98). ACM.
10. Kao, H. L. C., Ho, B. J., Lin, A. C., & Chu, H. H. (2012, September). Phone-based gait analysis to detect alcohol usage. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 661-662). ACM.
11. Fairley, J. A., Sejdić, E., & Chau, T. (2010). The effect of treadmill walking on the stride interval dynamics of children. *Human movement science*, *29*(6), 987-998.
12. Brach, J. S., McGurl, D., Wert, D., VanSwearingen, J. M., Perera, S., Cham, R., & Studenski, S. (2010). Validation of a measure of smoothness of walking. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, glq170.
13. Porta, A., Baselli, G., Liberati, D., Montano, N., Cogliati, C., Gnecchi-Ruscone, T., ... & Cerutti, S. (1998). Measuring regularity by means of a corrected conditional entropy in sympathetic outflow. *Biological cybernetics*, *78*(1), 71-78.
14. Porta, A., Guzzetti, S., Montano, N., Furlan, R., Pagani, M., Malliani, A., & Cerutti, S. (2001). Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series. *Biomedical Engineering, IEEE Transactions on*, *48*(11), 1282-1291.
15. Lee, J., Sejdic, E., Steele, C. M., & Chau, T. (2010). Effects of liquid stimuli on dual-axis swallowing accelerometry signals in a healthy population. *Biomed Eng Online*, *9*(7).
16. Rosso, O. A., Blanco, S., Yordanova, J., Kolev, V., Figliola, A., Schürmann, M., & Başar, E. (2001). Wavelet entropy: a new tool for analysis of short duration brain electrical signals. *Journal of neuroscience methods*, *105*(1), 65-75.
17. Aboy, M., Hornero, R., Abásolo, D., & Álvarez, D. (2006). Interpretation of the Lempel-Ziv complexity measure in the context of biomedical signal analysis. *Biomedical Engineering, IEEE Transactions on*, *53*(11), 2282-2288.
18. Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *Information Theory, IEEE Transactions on*, *22*(1), 75-81.
19. Ferenets, R., Lipping, T., Anier, A., Jäntti, V., Melto, S., & Hovilehto, S. (2006). Comparison of entropy and complexity measures for the assessment of depth of sedation. *Biomedical Engineering, IEEE Transactions on*, *53*(6), 1067-1077.
20. Demura, S., & Uchiyama, M. (2008). Influence of moderate alcohol ingestion on gait. *Sport Sciences for Health*, *4*(1-2), 21-26.
21. Room, R., Babor, T., & Rehm, J. (2005). Alcohol and public health. *The lancet*, *365*(9458), 519-530.

22. Izzi, M. (2014). SCRAM Bracelet Laws. Available: http://www.legalmatch.com/law-library/article/scram-bracelet-laws.html. [Access 04 April 2016].
23. Substance Abuse and Mental Health Services Administration (SAMHSA). 2013 National Survey on Drug Use and Health (NSDUH). Table 2.41B—Alcohol Use in Lifetime, Past Year, and Past Month among Persons Aged 18 or Older, by Demographic Characteristics: Percentages, 2012 and 2013. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabsPDFWHTML2013/Web/HTML/NSDUH-DetTabsSect2peTabs1to42-2013.htm#tab2.41b. [Access 04 April 2016].
24. SAMHSA. 2013 National Survey on Drug Use and Health (NSDUH). Table 2.46B—Alcohol Use, Binge Alcohol Use, and Heavy Alcohol Use in the Past Month among Persons Aged 18 or Older, by Demographic Characteristics: Percentages, 2012 and 2013. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabsPDFWHTML2013/Web/HTML/NSDUH-DetTabsSect2peTabs43to84-2013.htm#tab2.46b. [Access 04 April 2016].
25. Multiple Cause of Death Public-Use Data File, 2012 and 2013. AAFs from CDC Alcohol-Related Disease Impact (ARDI). National Survey on Drug Use and Health, 2012 and 2013 for estimating indirect AAFs for Liver Cancer. Available at: http://www.cdc.gov/alcohol/ardi.htm. [Access 04 April 2016].
26. National Cancer Institute. Cancer Trends Progress Report, 2011–2012 Update. Available at: http://progressreport.cancer.gov/sites/default/files/archive/report2011.pdf. [Access 04 April 2016].
27. Centers for Disease Control and Prevention. Alcohol use and health. Available at: http://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm. [Access 04 April 2016].
28. National Institute on Alcohol Abuse and Alcoholism (NIAAA). NIAAA Council Approves Definition of Binge Drinking. NIAAA Newsletter, Number 3, Winter 2004. Available at: http://pubs.niaaa.nih.gov/publications/Newsletter/winter2004/Newsletter_Number3.pdf. [Access 04 April 2016].
29. SCRAM Systems, "SCRAM Continuous Alcohol Monitoring," Alcohol Monitoring Systems, Inc., 2014. [Online]. Available: http://www.scramsystems.com/index/scram/continuous-alcoholmonitoring. [Accessed 09 December 2014].
30. Tokyoflash Japan, "Kisai Intoxicated LCD Watch," Tokyoflash Japan, 2014. [Online]. Available: http://www.tokyoflash.com/en/watches/kisai/intoxicated/. [Accessed 09 December 2014].
31. Myrecek, “AlcoDroid Alcohol Tracker” [Online]. Available: https://play.google.com/store/apps/details?id=org.M.alcodroid&hl=en. [Accessed 04 April 2016].
32. Raphael Wichmann, “IntelliDrink PRO – Blood Alcohol Content (BAC) Calculator”. [Online] Available: https://itunes.apple.com/us/app/intellidrink-pro-blood-alcohol/id440759306?mt=8. [Access 04 April 2016].
33. Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern* Analysis and Machine Intelligence, IEEE Transactions on, 11(7), 674-693.
34. Hall, M. A. (1999). *Correlation-based feature selection for machine learning* (Doctoral dissertation, The University of Waikato).
35. Alpaydin, E. (2014). *Introduction to machine learning*. MIT press. p. 9.
36. Ho, T. K. (1995, August). Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on* (Vol. 1, pp. 278-282). IEEE.
37. Friedman, J., Hastie, T., & Tibshirani, R. (2008). *The elements of statistical learning (2nd ed.).* Springer, Berlin: Springer series in statistics.
38. Salzberg, S. L. (1994). C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16(3), 235-240.
39. Umd.edu. Top 10 Algorithms in Data Mining. [Online] Available: http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf. [Accessed 04 April 2016]
40. Cohen, W. W. (1995, July). Fast effective rule induction. In *Proceedings of the twelfth international conference on machine learning* (pp. 115-123).
41. Russell, S., Norvig, P., & Intelligence, A. (2003) [1995]. *Artificial Intelligence: A modern approach (2nd ed.).* Prentice-Hall, Egnlewood Cliffs, 25, 27.
42. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.
43. Queensu.ca. Gait analysis offers a unique means to measure the mechanical factors of joint loading, orientation, and neuromuscular function during activities of daily living such as walking. [Online] Available: http://me.queensu.ca/People/Deluzio/Gait.html. [Accessed 04 April 2016]

44. Drunk Buster Goggles. *Drunk Busters of America, LLC*. [Online] Available: http://www.drunkbusters.com. [Accessed 04 April 2016]
45. Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering, 6*(5), 551.
46. S. P. Awate, "Maximum-a-Posteriori (MAP) Estimation," University of Utah, 21 February 2007. [Online]. Available: https://www.cs.utah.edu/~suyash/Dissertation_html/node8.html. [Accessed 25 March 2015].
47. Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.
48. Begg, R. K., Palaniswami, M., & Owen, B. (2005). Support vector machines for automated gait classification. Biomedical Engineering, IEEE Transactions on, 52(5), 828-838.
49. Kecman, V. (2001). *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. MIT press.
50. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.
51. Wallén, M. B., Nero, H., Franzén, E., & Hagströmer, M. (2014). Comparison of two accelerometer filter settings in individuals with Parkinson's disease. *Physiological measurement*, 35(11), 2287.
52. Madeleine, P., Tuker, K., Arendt-Nielsen, L., & Farina, D. (2007). Heterogeneous mechanomyographic absolute activation of paraspinal muscles assessed by a two-dimensional array during short and sustained contractions. *Journal of biomechanics*, 40(12), 2663-2671.
53. Montgomery Jr, E. B. (2006). *U.S. Patent No. 7,136,696*. Washington, DC: U.S. Patent and Trademark Office.
54. Salarian, A., Russmann, H., Vingerhoets, F. J., Dehollain, C., Blanc, Y., Burkhard, P. R., & Aminian, K. (2004). Gait assessment in Parkinson's disease: toward an ambulatory system for long-term monitoring. *Biomedical Engineering, IEEE Transactions* on, 51(8), 1434-1443.
55. Akay, M. (1992). Noninvasive diagnosis of coronary artery disease using a neural network algorithm. *Biological cybernetics*, 67(4), 361-367.
56. Pappas, I. P., Keller, T., Mangold, S., Popovic, M. R., Dietz, V., & Morari, M. (2004). A reliable gyroscope-based gait-phase detection sensor embedded in a shoe insole. *Sensors Journal, IEEE, 4*(2), 268-274.
57. Titianova, E. B., Mateev, P. S., & Tarkka, I. M. (2004). Footprint analysis of gait using a pressure sensor system. *Journal of Electromyography and Kinesiology, 14*(2), 275-281.
58. Henriksen, M., Lund, H., Moe-Nilssen, R., Bliddal, H., & Danneskiod-Samsøe, B. (2004). Test–retest reliability of trunk accelerometric gait analysis. *Gait & posture, 19*(3), 288-297.
59. Moe-Nilssen, R., & Helbostad, J. L. (2004). Estimation of gait cycle characteristics by trunk accelerometry. *Journal of biomechanics, 37*(1), 121-126.
60. Hartmann, A., Murer, K., de Bie, R. A., & de Bruin, E. D. (2009). Reproducibility of spatio-temporal gait parameters under different conditions in older adults using a trunk tri-axial accelerometer system. *Gait & posture, 30*(3), 351-355.
61. Munoz, M., Argoul, P., & Farges, F. (2003). Continuous Cauchy wavelet transform analyses of EXAFS spectra: A qualitative approach. *American mineralogist, 88*(4), 694-700.
62. *Cross-correlation*. [Online] Available: https://en.wikipedia.org/wiki/Cross-correlation [Accessed 04 April 2016]
63. *P-value*. [Online] Available: https://en.wikipedia.org/wiki/P-value [Accessed 04 April 2016]
64. Walker, J. (2005). *The Hacker's Diet* [Online]. Available: https://www.fourmilab.ch/hackdiet/e4/ [Accessed 04 April 2016]

Appendix A: Data Samples

Table 37 Data Sample, One Person One Group of 4 Segments. Each Segment is sampled with an approximate frequency of 10Hz, covering a total time of 5 seconds. This group of data is related to BAC=0.

Accelerometer x (m/s^2)	Accelerometer y (m/s^2)	Accelerometer z (m/s^2)	Time stamp (s)
0.68354	-8.6592	-1.9279	0
1.4389	-5.7048	0.89364	0.09
120	-9.7743	-0.32082	0.189
48	-9.3374	-3.225	0.288
0	-9.3643	1.2821	0.388
5.7419	-13.969	2.7545	0.487
-4.5514	-2.8383	-1.5185	0.586
2.2116	-7.2197	-1.0313	0.685
-0.6608	-8.8149	-2.1805	0.784
0.18316	-13.743	-4.8836	0.883
2.2248	-5.7718	0.9517	0.982
1.5473	-9.7073	-1.7711	1.081
1.5227	-5.7569	1.3162	1.18
-0.56324	-10.655	-0.70449	1.279
0.098162	-11.241	-1.479	1.378
1.549	-11.808	5.3534	1.477
-0.58359	-16.654	1.4982	1.576
0.50578	-5.4779	-0.06045	1.675
0.73203	-7.6596	-2.0734	1.774
-1.5101	-10.862	-2.2817	1.873
1.2857	-17.183	0.68055	1.972
1.0223	-12.392	1.4856	2.071
1.4892	-6.0603	-2.7599	2.17
-1.6215	-10.491	-0.72664	2.269
0.07841	-8.8753	-0.85593	2.369
-1.3473	-12.253	-4.2066	2.482
-1.3216	-12.566	1.2983	2.582
-1.3216	-12.566	1.2983	2.667
1.2061	-4.9021	-1.0145	2.766
-0.2466	-11.314	-1.8417	2.865
-1.1756	-10.37	-3.7667	2.965
2.8784	-7.8835	3.2316	3.063
2.7988	-10.991	-0.02634	3.163
1.7615	-5.3852	1.1899	3.262
-0.65841	-10.633	-1.0463	3.362
-0.0826	-10.705	-1.6059	3.461
-0.90261	-9.0519	2.7186	3.56
1.8824	-18.872	0.083797	3.659

-1.2522	-4.1677	-0.26635	3.758
2.1817	-7.8703	-2.2075	3.857
-1.4311	-9.0465	-2.6348	3.956
-1.0732	-12.999	-4.3964	4.056
2.2948	-6.672	1.2947	4.155
2.077	-9.6349	-1.1259	4.254
0.8667	-7.2951	0.83617	4.353
-1.497	-9.3984	-0.50578	4.452
1.6334	-10.169	-1.9207	4.551
-2.0632	-10.411	-2.7013	24.84
1.3306	-17.047	0.62848	24.929
-0.9068	-10.998	-0.246	25.028
2.1955	-6.0831	-2.2212	25.127
-1.6873	-10.572	0.28192	25.226
-0.66499	-6.5697	-2.955	25.326
-3.7356	-10.8	-0.27054	25.425
4.3802	-8.2893	5.5593	25.524
-3.8571	-9.0363	0.85293	25.623
1.6879	-6.3596	-1.0062	25.722
-0.99419	-10.215	-1.8735	25.821
-0.31544	-13.305	-5.36	25.92
7.0252	-10.335	3.2896	26.019
4.3365	-12.743	1.3563	26.118
0.7452	-6.5038	0.34058	26.217
-2.2553	-9.6432	-2.0147	26.316
-0.67756	-9.5942	-0.85234	26.416
-0.83797	-7.683	0.33758	26.515
4.2341	-18.372	2.8174	26.614
-3.8589	-2.885	-1.4329	26.713
2.3804	-8.1828	-1.7472	26.812
-1.9896	-8.6664	-2.8503	26.911
-1.3976	-13.27	-3.486	27.01
-1.3803	-7.3101	-0.7841	27.109
5.8407	-9.8617	-2.2045	27.208
-0.5351	-9.7318	1.257	27.307
-0.90202	-8.9142	-1.2151	27.406
-2.1793	-13.975	-2.3403	27.506
3.699	-8.767	5.5701	27.605
-3.7948	-12.126	2.0093	27.704
1.4078	-5.3732	-0.94511	27.803
-0.4118	-8.8424	-2.0985	27.902
-0.33818	-13.642	-4.3616	28.001
6.2106	-7.2754	1.9225	28.1
0.56503	-12.122	0.7027	28.199
0.91758	-4.4748	0.39325	28.298
-1.4251	-11.955	-0.8685	28.397

-0.81642	-7.7069	-1.6604	28.496
-1.3886	-8.488	1.1558	28.596
4.2359	-17.617	1.9082	28.695
-3.5434	-1.8046	-1.6538	28.794
2.9838	-8.433	-1.5006	28.893
-2.5373	-8.8053	-3.1328	28.993
0.61292	-16.814	-0.64105	29.092
-1.2408	-10.488	0.37769	29.191
3.2471	-5.2816	-3.7379	29.291
-2.8635	-10.959	0.48542	29.39
-0.3699	-7.437	-2.5965	29.49
-2.6839	-9.988	-1.8328	29.589
6.6283	-16.225	6.3901	29.688
-6.6738	-1.1462	-2.9072	29.787
2.3397	-7.3113	-1.2031	29.887
-3.0281	-8.6395	-2.7623	29.986
-0.77273	-11.873	-4.5657	30.085
-0.0012	-6.4775	1.5473	30.185
3.5303	-10.032	-1.506	30.284
0.011372	-7.9805	1.0409	30.383
-0.51535	-9.8402	-1.2306	30.482
-1.5449	-12.715	-3.45	30.581
1.8777	-8.9172	5.5767	30.68
-2.2601	-15.725	1.0241	30.779
0.38547	-4.8357	-1.1612	30.879
0.83438	-8.7382	-1.7334	30.978
-2.8772	-13.521	-3.3046	31.077
5.3056	-17.253	3.8972	31.176
0.18555	-15.025	0.69372	31.275
-0.19872	-4.6394	0.067038	31.375
-0.97564	-11.522	-1.1678	31.474
-0.36811	-8.19	-2.1081	31.573
-1.8112	-10.383	3.5925	31.672
2.9981	-19.613	0.55246	31.771
-2.0943	-4.3574	0.58658	31.87
3.3471	-7.5268	-2.627	31.969
-3.2154	-11.024	-3.1406	32.068
0.2873	-12.97	-0.45789	32.167
-1.1235	-10.497	-0.75058	32.266
3.6859	-7.1353	-2.6941	32.365
-0.67876	-8.7951	0.43934	32.464
-1.0529	-7.3568	-1.9315	32.563
-2.6558	-11.99	-0.99958	32.662
5.2834	-11.325	6.6924	32.762
-5.0003	-6.8211	-0.60454	32.861
2.2452	-7.3251	-1.825	32.96

-0.62728	-8.9082	-1.5862	33.059
0.080206	-12.813	-4.7285	33.158
1.8902	-6.3776	2.2188	33.258
3.2711	-10.694	-1.309	33.357
0.7003	-6.2782	0.89124	33.456
-1.7741	-9.8647	-1.3114	33.555
0.31663	-12.252	-2.3655	33.654
-1.2594	-9.7707	2.7037	33.753
2.5271	-19.613	-0.09038	33.852
-0.64224	-4.0606	0.33339	33.951
2.0668	-8.7251	-1.3066	34.05
-2.2075	-6.3321	-3.3549	34.149
-2.5792	-12.526	-3.0783	34.248
0.36631	-8.8939	0.25857	34.347
3.2585	-8.9561	-1.7939	34.446
-2.0859	-6.702	0.7853	34.545
-0.84156	-10.968	-1.0068	34.644
-0.22805	-12.459	-3.2286	34.743
-0.13647	-8.4839	5.0578	34.842
2.7815	-8.2343	-2.7054	45.401
1.4593	-10.311	-1.6376	45.5
1.2216	-8.5964	-3.6919	45.599
0.72784	-10.97	4.4454	45.698
-1.0121	-17.249	0.29628	45.798
-0.03472	-4.5999	-1.9076	45.898
2.0027	-8.7179	-2.5169	45.997
-0.82959	-9.5469	-2.7707	46.096
-3.0801	-11.979	-3.1238	46.196
1.8064	-8.0637	-0.99599	46.295
2.0189	-14.014	1.1073	46.394
1.6197	-6.3225	-1.4712	46.494
-2.2482	-8.5682	-0.73981	46.593
0.35973	-8.84	-2.3308	46.692
-1.7184	-12.559	-3.1304	46.791
1.7765	-7.0988	5.5049	46.89
-1.7921	-13.026	2.745	46.989
0.5764	-5.1589	-1.6514	47.088
1.1091	-10.102	-2.5223	47.187
-3.9217	-10.615	-3.3387	47.286
3.1627	-17.502	2.651	47.385
0.11971	-12.728	1.0672	47.484
1.7262	-4.2096	-1.0654	47.583
-3.0029	-11.293	-0.80864	47.682
-1.5874	-6.1088	-3.1143	47.781
-2.4032	-9.1261	-0.39265	47.88
4.0881	-16.057	3.887	47.979

-4.5783	-2.5361	-1.8741	48.078
2.7551	-7.6638	-1.4066	48.178
-1.8525	-8.679	-2.742	48.277
-0.96067	-12.397	-4.7968	48.376
-0.6195	-11.168	1.2372	48.476
4.5334	-8.9992	-2.0836	48.575
-0.94751	-9.6917	1.2617	48.674
-2.0063	-8.357	-3.3818	48.773
-2.1051	-11.578	-3.7757	48.873
1.4676	-6.9677	5.6778	48.972
-3.7421	-14.356	-0.331	49.071
2.1997	-4.6214	-1.3635	49.17
1.2863	-8.5365	-2.873	49.269

Appendix B: Code Samples

```
%% Statistical Analysis on Features
…
%% Calculate Features
% load data and perform segmentation (5 sec)
k = 0;
segmentNum = 1;
for a=1:length(fileNames)
   % person a
   filename = char(fileNames(a));
   [data_x, data_y, data_z, data_t, data_r, data_s] = loadData_new(filename); % load one data file
   startIndex = k+1;
   % read by segments
   while(~isempty(data_s)&&data_r(1)<=60)
      j = sum(data_s <= segmentNum);
      x = data_x(1:j);
      y = data_y(1:j);
      z = data_z(1:j);
      t = data_t(1:j);
      k = k + 1;
      segmentNum = segmentNum+1;
      numDrink(k,1) = BACValue(data_r(1));
      % remove these lines
      data_x(1:j) = [];
      data_y(1:j) = [];
      data_z(1:j) = [];
      data_t(1:j) = [];
      data_r(1:j) = [];
      data_s(1:j) = [];
      % extract features
      if(j<=1)
         k = k - 1;
         continue;
      end
      x_o = x;
      y_o = y;
      z_o = z;
      x = denoise(x);
      y = denoise(y);
      z = denoise(z);
      [feature(k,1), loc] = numSteps(x, y, z);
% remove those with few steps
      if(feature(k,1)<=2)
         k = k - 1;
         continue;
      end
% calculation features
```

```
        feature(k,2) = averageStepTime(t, loc);
        feature(k,3) = averageCadence(t, loc);
        feature(k,4) = skewness_acc(x, y, z);
        feature(k,5) = kurtosis_acc(x, y, z);
        feature(k,6) = minMaxDiff(x, y, z);
        feature(k,7) = std_acc(x, y, z);
        feature(k,8) = rms_acc(x, y, z);
        feature(k,9) = coef_var_stepTime(t, loc);
        feature(k,10) = harmonicR(x, y, z);
        feature(k,11) = cross_corr(x, y, z);
        feature(k,12) = entropy_rate(x, y, z);
        feature(k,13) = averagePower(x, y, z);
        feature(k,14) = radioSpectralPeak(x, y, z);
        feature(k,15) = snr_acc(x_o, y_o, z_o);
        feature(k,16) = thd_acc(x_o, y_o, z_o);
        feature(k,17) = powerFreq_05_3(x, y, z);
        feature(k,18) = powerFreq_05_3_windowed(x, y, z);
        feature(k,19) = peakFreq(x, y, z);
        feature(k,20) = spectralCentroid(x, y, z);
        feature(k,21) = acc_bw(x, y, z);
        feature(k,22) = wavelet_band(x, y, z);
        feature(k,23) = wavelet_entropy(x, y, z);
        feature(k,24) = radioSpectralPeak_FFT(x, y, z);
        feature(k,25) = radioSpectralPeak_DCT(x, y, z);
        feature(k,26) = averageStepLength(t, loc);
        feature(k,27) = gaitVelocity(t, loc);
    end
    endIndex = k;
    if(startIndex==endIndex)
        feature(k,:) = [];
        numDrink(k,:) = [];
        k = k - 1;
        continue;
    end
    % normalization for person a
    % norm(k) = observation(segment k, drunkLevel i, person a)/avg(drunkLevel i, person a)
    featureMean = repmat(mean(feature(startIndex:endIndex,:), 1), endIndex-startIndex+1, 1);
end

%% Plot Feature-Level Distribution
% plot _feature_ vs _numDrink_
for i = 1:numFeatures
    figure;
    hold on;
    % Boxplot
    plot(numDrink+0.1, feature(:,i), '*');
    boxplot(feature(:,i), numDrink,...
        'Label', [0, 0.05, 0.12, 0.2, 0.3],...
```

```
      'Position', [0.1, 0.15, 0.22, 0.3, 0.4]);
   title(char(featureNames(i)));
   xlim([0 0.5]);
   set(gcf,'PaperUnits','inches','PaperPosition',[0 0 3 7])
   hold off;
   print(strcat('report\png_', num2str(i), '_', char(featureNames(i)), '__before'),'-dpng');
   close all;
end

%% Correlation
[R, P] = corrcoef([numDrink feature]);

%% CSV_Report
…
```

Appendix C: Normalization Results

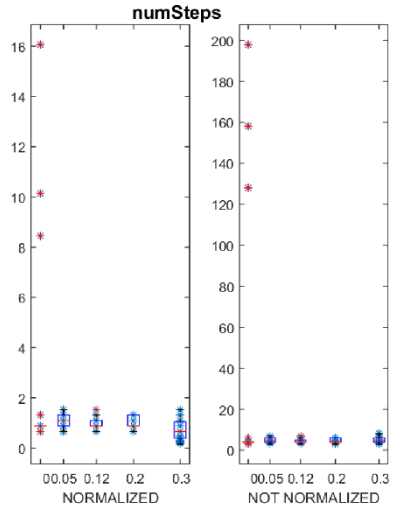

Figure 28 Data Distribution of Feature "Number of Steps" (Normalized on left vs. Not Normalized on right)

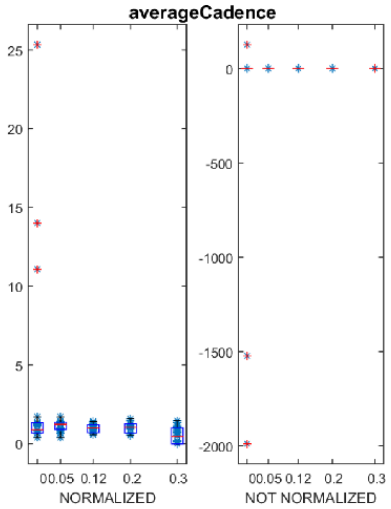

Figure 30 Data Distribution of Feature "Average Cadence" (Normalized on left vs. Not Normalized on right)

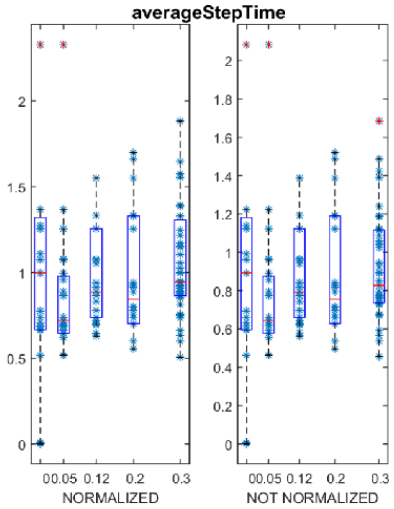

Figure 29 Data Distribution of Feature "Average Step Time" (Normalized on left vs. Not Normalized on right)

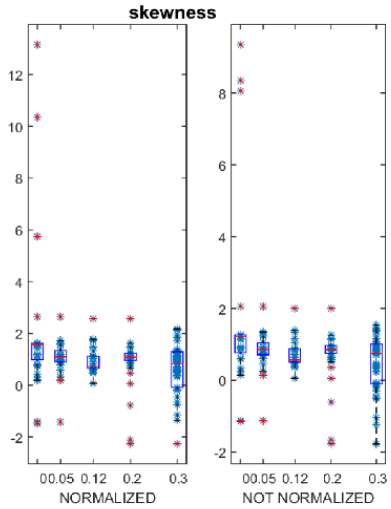

Figure 31 Data Distribution of Feature "Skewness" (Normalized on left vs. Not Normalized on right)

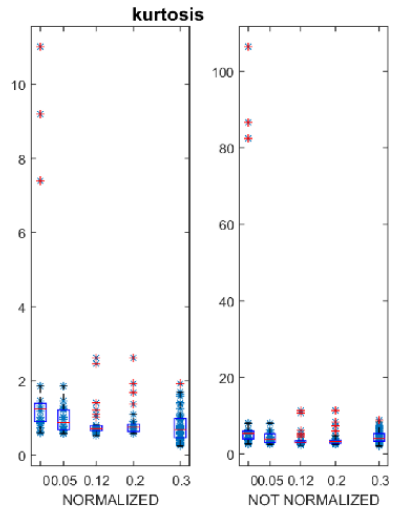

Figure 32 Data Distribution of Feature "Kurtosis" (Normalized on left vs. Not Normalized on right)

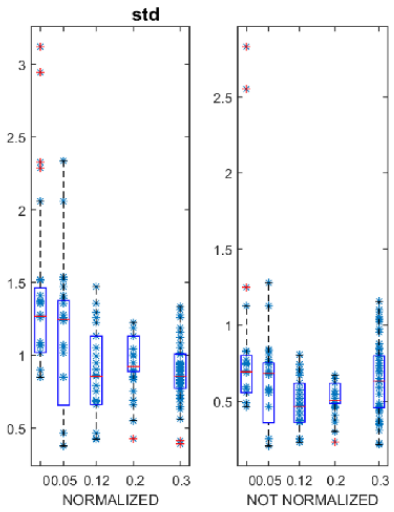

Figure 33 Data Distribution of Feature "Standard Deviation" (Normalized on left vs. Not Normalized on right)

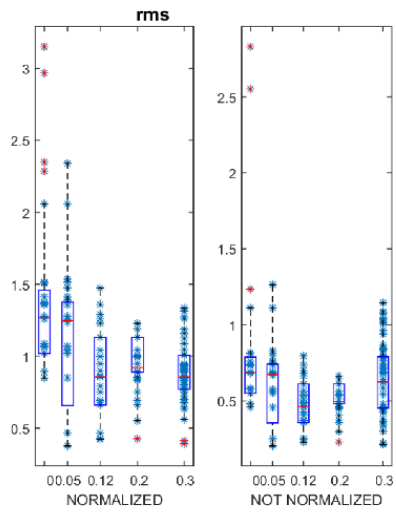

Figure 34 Data Distribution of Feature "Root Mean Square" (Normalized on left vs. Not Normalized on right)

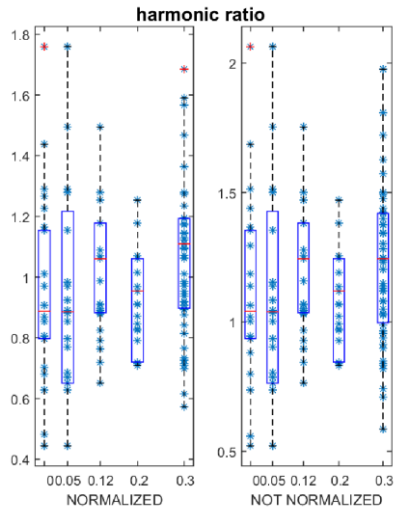

Figure 35 Data Distribution of Feature "Harmonic Ratio" (Normalized on left vs. Not Normalized on right)

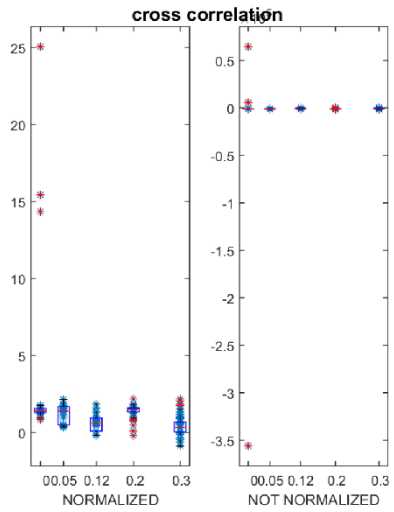

Figure 36 Data Distribution of Feature "Zeroth-Lag Cross-Correlation Coefficient" (Normalized on left vs. Not Normalized on right)

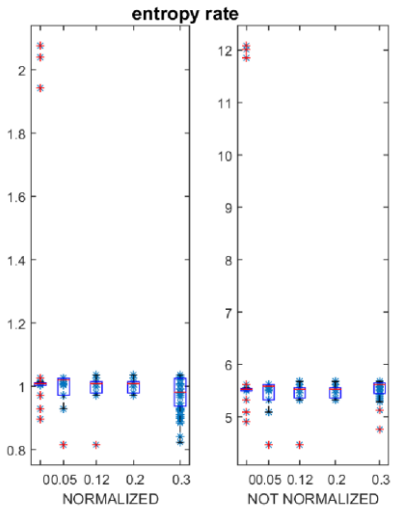

Figure 37 Data Distribution of Feature "Entropy Rate" (Normalized on left vs. Not Normalized on right)

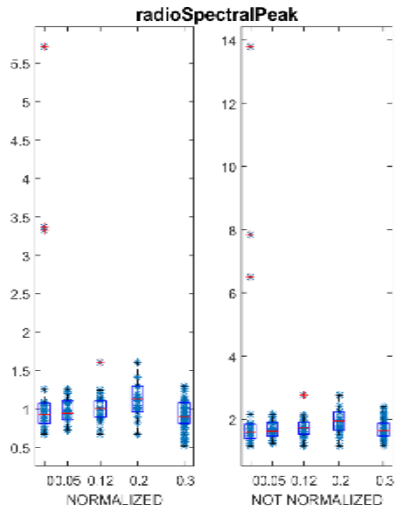

Figure 38 Data Distribution of Feature "Ratio of Spectral Peak" (Normalized on left vs. Not Normalized on right)

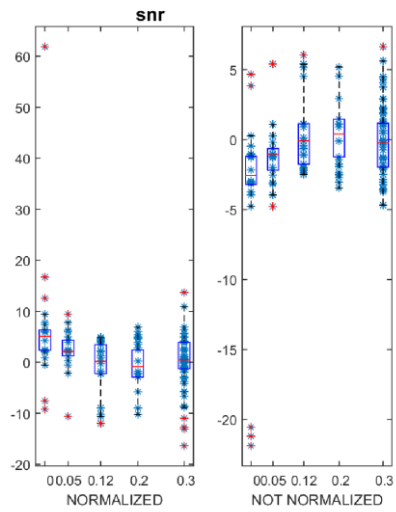

Figure 39 Data Distribution of Feature "Signal Noise Ratio" (Normalized on left vs. Not Normalized on right)

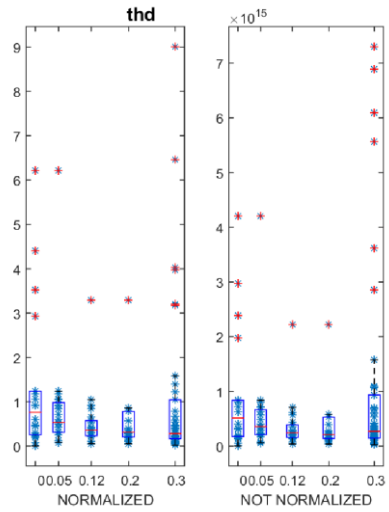

Figure 40 Data Distribution of Feature "Total Harmonic Distortion" (Normalized on left vs. Not Normalized on right)

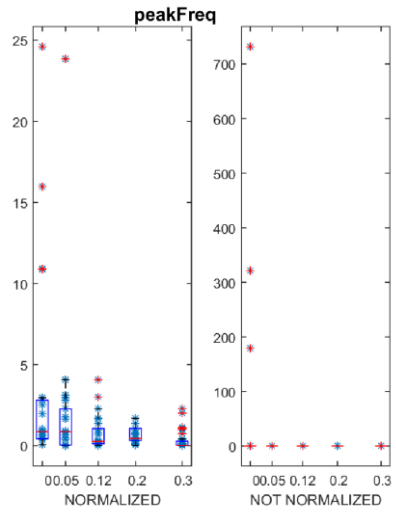

Figure 42 Data Distribution of Feature "Peak Frequency" (Normalized on left vs. Not Normalized on right)

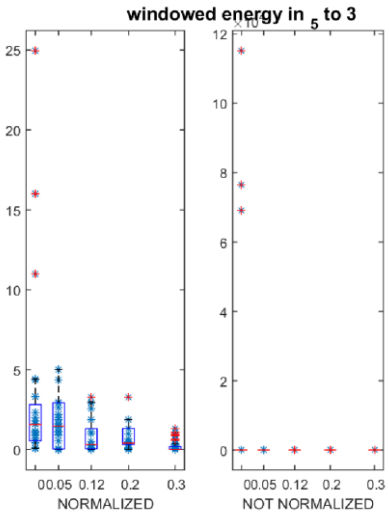

Figure 41 Data Distribution of Feature "Windowed Energy in Band 0.5 to 3 Hz" (Normalized on left vs. Not Normalized on right)

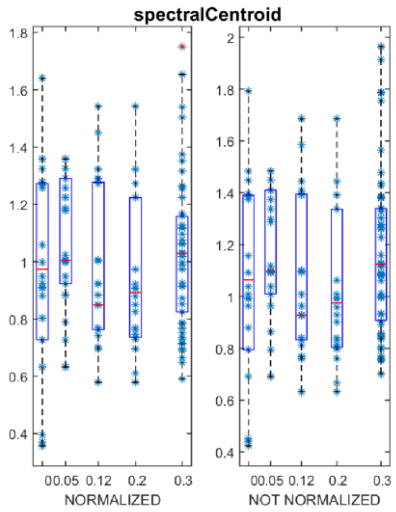

Figure 43 Data Distribution of Feature "Spectral Centroid" (Normalized on left vs. Not Normalized on right)

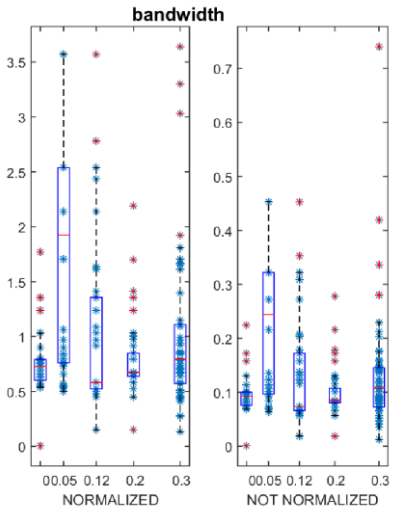

Figure 44 Data Distribution of Feature "Bandwidth" (Normalized on left vs. Not Normalized on right)

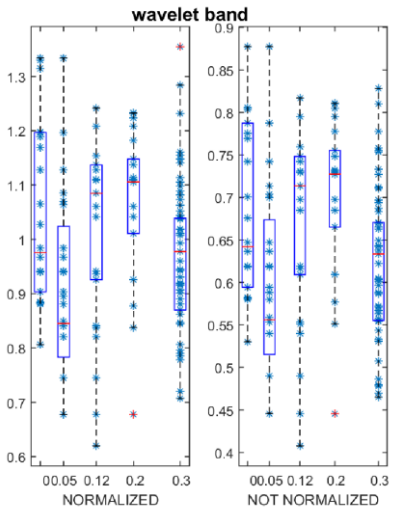

Figure 45 Data Distribution of Feature "Wavelet Bandwidth" (Normalized on left vs. Not Normalized on right)

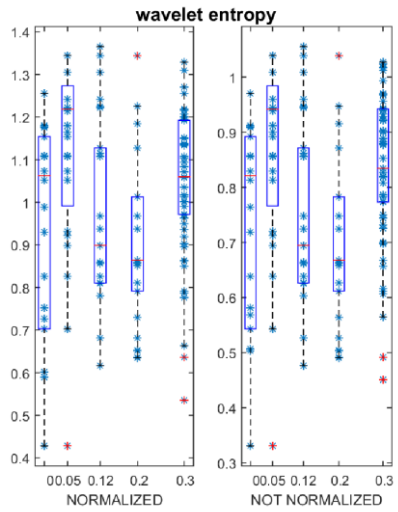

Figure 46 Data Distribution of Feature "Wavelet Entropy Rate" (Normalized on left vs. Not Normalized on right)

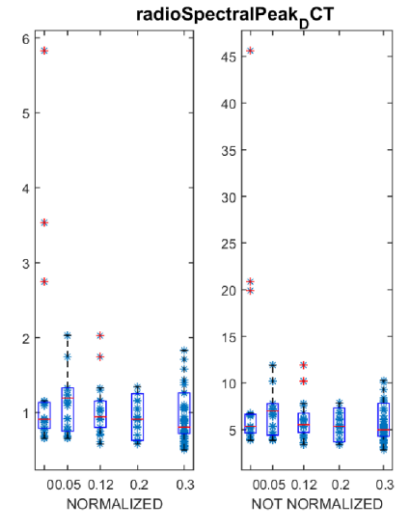

Figure 47 Data Distribution of Feature "Ratio of Spectral Peak by DCT" (Normalized on left vs. Not Normalized on right)

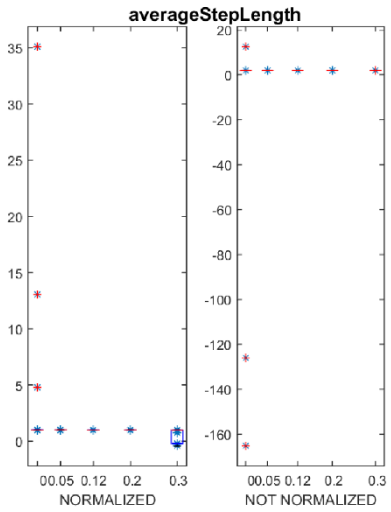

Figure 48 Data Distribution of Feature "Average Step Length" (Normalized on left vs. Not Normalized on right)

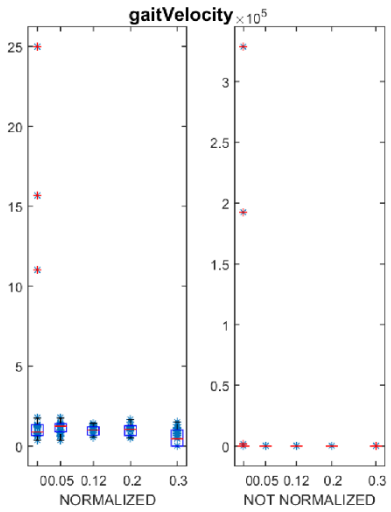

Figure 49 Data Distribution of Feature "Gait Velocity" (Normalized on left vs. Not Normalized on right)